Docket No. 1506.1002 (JDH)

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of:)
)
Yasuyuki FUJIKAWA)
) Group Art Unit: Unassigned
Serial No.: To be assigned)
) Examiner: Unassigned
Filed: January 4, 2001)

For: **STRUCTURAL DOCUMENTATION SYSTEM**

SUBMISSION OF CERTIFIED COPY OF PRIOR FOREIGN APPLICATION IN ACCORDANCE WITH THE REQUIREMENTS OF 37 C.F.R. §1.55

Assistant Commissioner for Patents
Washington, D.C. 20231

Sir:

In accordance with the provisions of 37 C.F.R. §1.55, the applicant submits herewith a certified copy of the following foreign application:

Japanese Patent Application No. 2000-027460
Filed: February 4, 2000.

It is respectfully requested that the applicant be given the benefit of the foreign filing date as evidenced by the certified papers attached hereto, in accordance with the requirements of 35 U.S.C. §119.

Respectfully submitted,

STAAS & HALSEY LLP

By: _____
James D. Halsey, Jr.
Registration No. 22,729

700 11th Street, N.W., Ste. 500
Washington, D.C. 20001
(202) 434-1500
Date: January 4, 2001

日 本 国 特 許 庁
PATENT OFFICE
JAPANESE GOVERNMENT

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。
This is to certify that the annexed is a true copy of the following application as filed with this Office.

CERTIFIED COPY OF PRIORITY DOCUMENT

出願年月日
Date of Application: ２０００年 ２月 ４日

出願番号
Application Number: 特願２０００－０２７４６０

出願人
Applicant (s): 富士通株式会社

２０００年 ８月１８日

特許庁長官
Commissioner,
Patent Office 及川耕造

出証番号 出証特２０００－３０６４８９４

【書類名】　特許願
【整理番号】　9903156
【あて先】　特許庁長官殿
【国際特許分類】　G06F 13/368
【発明者】
【住所又は居所】　神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内
【氏名】　藤川　泰之
【特許出願人】
【識別番号】　000005223
【氏名又は名称】　富士通株式会社
【代理人】
【識別番号】　100098235
【弁理士】
【氏名又は名称】　金井　英幸
【手数料の表示】
【予納台帳番号】　062606
【納付金額】　21,000円
【提出物件の目録】
【物件名】　明細書　1
【物件名】　図面　1
【物件名】　要約書　1
【包括委任状番号】　9908696
【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　構造文書化システム

【特許請求の範囲】

【請求項１】

テキスト形式で記述された処理対象の電子文書を、所定の文書構造を有する構造化文書に変換するための構造文書化システムであって、

前記文書構造を構成する基本単位である各要素間の相互関係を定義するとともに各要素毎にその抽出条件及び識別子を定義した定義情報を読み込む読込部と、

前記読込部によって読み込まれた定義情報によって定義された各要素毎の抽出条件を順次参照し、参照した要素の抽出条件に合致した領域を前記処理対象の電子文書から抽出する検索部と、

前記検索部によって各要素に関して抽出された領域を、前記定義情報によって定義された各要素間の相互関係に従って組み合わせるとともに、各領域に対して前記定義情報によって定義された識別子を付すことによって前記構造化文書を生成する構造化文書生成部と

を備えたことを特徴とする構造文書化システム。

【請求項２】

前記構造化文書生成部は、前記識別子として、前記検索部によって抽出された各領域の前後にタグを付す

ことを特徴とする請求項１記載の構造文書化システム。

【請求項３】

前記定義情報によって定義された各要素間の相互関係には、一つの上位階層の要素内に複数の下位階層の要素が含まれてなる階層構造が含まれ、

前記検索部は、前記上位階層の抽出条件に合致して抽出した領域に対して、前記各下位階層の要素の抽出条件を参照した抽出を行い、

前記構造化文書生成部は、下位階層を持たない末端の要素については前記検索部によって抽出された領域の前後にタグを付し、下位階層を持つ要素については当該要素の下位階層の全要素について前記検索部によって夫々抽出された領域を合わせてなる領域の前後にタグを付す

ことを特徴とする請求項２記載の構造文書化システム。

【請求項４】

前記定義情報によって定義された各要素間の相互関係には、一つの上位階層の要素内に繰り返し構造を持つ下位階層の要素が含まれてなる階層構造が含まれ、

前記検索部は、前記上位階層の抽出条件に合致して抽出した領域に対して、繰り返し構造を持つ下位階層の要素の抽出条件を参照した抽出を、この抽出条件に合致する領域が抽出できなくなるまで繰り返し行い、

前記構造化文書生成部は、前記繰り返し構造を持つ下位階層の要素については前記検索部によって抽出された各領域の前後に夫々共通のタグを付す

ことを特徴とする請求項３記載の構造文書化システム。

【請求項５】

前記定義情報によって定義された各要素間の相互関係には、一つの上位階層の要素内に抽出順序制約を持つ要素を含む複数の下位階層の要素が含まれてなる階層構造が含まれ、

前記検索部は、前記上位階層の抽出条件に合致して抽出した領域内において、前記抽出順序制約を持つ下位階層の要素の抽出条件を参照した抽出を、他の下位階層の要素の抽出条件に合致して既に抽出された領域の直後から行う

ことを特徴とする請求項３記載の構造文書化システム。

【請求項６】

前記定義情報によって定義された何れかの要素の抽出条件は、抽出すべき領域全体の記述パターンである

ことを特徴とする請求項１記載の構造文書化システム。

【請求項７】

前記定義情報によって定義された何れかの要素の抽出条件は、抽出すべき領域の開始部分の記述パターン及び終了部分の記述パターンである

ことを特徴とする請求項１記載の構造文書化システム。

【請求項８】

前記記述パターンは、抽出すべき領域の文字列によって表現されている

ことを特徴とする請求項６又は７記載の構造文書化システム。

【請求項９】

前記記述パターンは、抽出すべき領域の文字列に対応した正規表現によって表現されている

ことを特徴とする請求項６又は７記載の構造文書化システム。

【請求項１０】

前記定義情報によって定義された何れかの要素の抽出条件は、抽出すべき領域の構文要素である

ことを特徴とする請求項１記載の構造文書化システム。

【請求項１１】

コンピュータに対して、

テキスト形式で記述された処理対象の電子文書を読み込ませ、

構造化文書の文書構造を構成する基本単位である各要素間の相互関係を定義するとともに各要素毎にその抽出条件及び識別子を定義した定義情報を読み込ませ、

読み込んだ定義情報によって定義された各要素毎の抽出条件を順次参照させ、

参照した要素の抽出条件に合致した領域を前記処理対象の電子文書から抽出させ、

各要素の抽出条件に合致するものとして抽出された領域を、前記定義情報によって定義された各要素間の相互関係に従って組み合わさせるとともに、各領域に対して前記定義情報によって定義された識別子を付加させることによって前記構造化文書を生成させる

プログラムを格納したコンピュータ可読媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、テキストやソースプログラムリスト等の電子文書から、予め定義された所定の構造を有する領域を自動的に識別し、識別した領域に当該構造に対応したタグを付加することによって構造化文書に変換する構造文書化システムに、関する。

【０００２】

【従来の技術】

構造化文書とは、テキスト形式で記述された一般文書やソースプログラムリスト等の電子文書に、その電子文書内の各領域の意義（例えば，その領域の記述内容が当該電子文書のヘッダ部であるという事，その領域の記述内容が当該電子文書の作成日時を示す，その領域の記述内容が当該電子文書の作成者名であるという事，その領域の記述内容が閲覧ソフト上において拡大表示されるという事，等）を示すタグを付加したものである。その構造化文書には、そのタグの付加規約に従って、ＸＭＬ，ＳＧＭＬ，ＨＴＭＬといった形式がある。このうちＸＭＬ，ＳＧＭＬは、タグの種類をユーザが任意に設定できるものであり、ＸＭＬは、ＳＧＭＬよりもその自由度を大きくしたものである。このような構造化文書において、各タグが付された領域相互の関係によって定義される電子文書の構造（例えば、ヘッダの次に本文が続き、ヘッダ内はタイトル，作成者，作成日時から構成される…といった構造）は、ＤＴＤ（文書構造定義）と、呼ばれる。

【０００３】

図３３（ａ）は、ＸＭＬによる構造化文書の一例を示し、同図（ｂ）は、このＸＭＬ文書のＤＴＤ（文書構造定義）をツリー形式に表した図である。これら両図を照らし合わせることによって理解されるように、ＤＴＤにおいて、構造化文章を構成する要素（一連の意義を持った領域）は、階層構造を有しており、それぞれに要素名（図３３（ｂ）においては「報告書」，「ヘッダ」，「タイトル」…）が与えられている。即ち、最上位階層の要素「報告書」は、文書全体であって、要素「ヘッダ」及び複数の要素「内容」から構成されている。さらに、要素「ヘッダ」は、要素「タイトル」，要素「日付」，要素「担当者」及び要素「顧客名」から構成されている。そして、構造化文書のテキスト中における各要素の前後には、図３３（ｂ）に示されるように各要素の要素名に対応したタグが付されている。例えば、要素「日付」の領域は、この要素名「日付」に対応したタグ＜DATA＞～＜/DATA＞によって囲まれている。従って、ＸＭＬまたはＳＧＭＬ文書を扱うシステム（以下、「ＸＭＬ／ＳＧＭＬシステム」という）は、このタグに囲まれている要素“1998.02.17”が日付を示しているものと、認識する。なお

、図３３においては、説明を容易にするために、各要素名を日本語にて「報告書」，「ヘッダ」，「タイトル」…と表したが、実際には、ＸＭＬ及びＳＧＭＬの規約に従い、タグと同じアルファベットで表された名称となっていなければならない。

【０００４】

このような構造化文書は、バイナリーファイルと異なり、基本的にテキストファイルであるが故にアプリケーションに依存しないという利点があるために、近年のインターネットの普及を背景に、インターネット等を通じて情報を交換したりデータベースにて管理するための文書形式として、広く用いられるようになって来ている。従って、このような構造化文書が定着する以前において作成された膨大な数量の電子文書を構造化文書に変換して、統一的に取り扱えるようにするという要請がある。従来、このような既存の電子文書の構造化文書への変換は、オペレータがエディタ画面上でその電子文書の内容を吟味して、ＤＴＤを参照しつつ、電子文書の意味内容に適したタグを手入力によって付加するしかなかった。

【０００５】

一方、構造化文書の対象となる電子文書であるプログラムソースに関しては、従来から、ＢＮＦ（バッカス・ナウア・フォーム）に従った構文要素とコメントとの両方を解析して必要な情報を抽出するツールが、出始めている。しかしながら、従来のツールは、抽出できる内容や出力形式が固定であって、柔軟性がなかった。

【０００６】

【発明が解決しようとする課題】

本発明の課題は、以上のような現状に鑑みて、テキスト形式で記述された処理対象電子文書に基づいて自動的に構造化文書を生成する構造文書化システムの提供である。

【０００７】

【課題を解決するための手段】

上記課題を解決するために、本発明による構造文書システムは、テキスト形式

で記述された処理対象の電子文書を所定の文書構造を有する構造化文書に変換するために、前記文書構造を構成する基本単位である各要素間の相互関係を定義するとともに各要素毎にその抽出条件及び識別子を定義した定義情報を読み込む読込部と、前記読込部によって読み込まれた定義情報によって定義された各要素毎の抽出条件を順次参照し、参照した要素の抽出条件に合致した領域を前記処理対象の電子文書から抽出する検索部と、前記検索部によって各要素に関して抽出された領域を、前記定義情報によって定義された各要素間の相互関係に従って組み合わせるとともに、各領域に対して前記定義情報によって定義された識別子を付すことによって前記構造化文書を生成する構造化文書生成部とを、備えたことを特徴とする。

【０００８】

このように構成された本発明による構造文書化システムによると、読込部によって読み込まれる定義情報には、変換の目標である構造化文書の文書構造を構成する各要素間の相互関係，及び、各要素に付されるべき識別子の他、各要素に対応する領域を処理対象電子文書から抽出する際の抽出条件が、定義されている。従って、検索部は、各要素毎にその抽出条件を順次参照することにより、各要素の抽出条件に合致した領域を、処理対象電子文書から抽出することができる。その結果、構造化文書生成部は、検索部によって抽出された領域を、定義情報によって定義された要素間の相互関係に従って組み合わせるとともに、各領域に対してその領域に対応した要素について定義情報によって定義された識別子を付すことによって、構造化文書として生成することができる。

【０００９】

本発明において処理対象とされる電子文書は、テキスト形式で記述されていさえすれば良いので、一般的な文章の他、Ｊａｖａソース等のソースプログラムリストも、これに含まれる。なお、ソースプログラムリスト内には、一般的な文章であるコメントが含まれていても良い。本発明において変換対象となる構造化文書，即ち、定義情報によって定義された識別子の種類は、ＸＭＬ形式によるものであっても良いし、ＳＧＭＬ形式によるものであっても良い。これら各形式によるものである場合には、識別子は、各領域の前後に付されるタグである。

【００１０】

【発明の実施の形態】

以下、本発明による構造文書化システムの実施の形態を、図面を参照して説明する。

（実施形態の概略）

本発明による構造文書化システムは、図２に示すように互いにバスＢによって接続されたＣＰＵ１，ハードディスク２，ＲＡＭ３，ディスプレイ４及び入力装置８から構成される一般的なコンピュータシステムにおいて、ハードディスク２に格納されたプログラムを、ＣＰＵ１によってＲＡＭ３上に読み出して、入力装置（キーボード及びマウス）８を介して入力されるオペレータからの操作に従って順次ＣＰＵ１によって実行して、その処理結果をディスプレイ４上に表示することによって、実現される。即ち、ハードディスク２が、本発明におけるコンピュータ可読媒体に該当し、ＣＰＵ１及びＲＡＭ３が、本発明における読込部，検索部，構造化文書生成部，コンピュータに、該当する。なお、図２においては、文章構造化システムを構成する全ハードウェアをローカルコンピュータのものとして示したが、この構造文書化システムは、複数のコンピュータをＬＡＮやインターネット等のネットワークを介して接続させてなる分散処理システムとして実現されても良い。

【００１１】

次に、以上のようにして実現される構造文書化システムの概略を説明する。図１は、本実施形態による構造文書化システム５の概念を、その拡張機能であるＤＴＤ＋パターン編集システム６及びＤＴＤ＋パターン作成支援システム７の概念と共に示した概念図である。同図に示されるように、構造文書化システム５は、テキスト形式で記述された一般文書やＢＮＦによる文法に従って記述されたソースプログラムリスト等を、処理対象文書Ｔとする。また、この構造文書化システム５は、最終的に生成すべき構造化文書の文章構造（ＤＴＤ）を構成する各要素間の相互関係を定義するとともに、この文章構造中の各要素に対応した領域を処理対象文書Ｔから自動的に抽出する際のキーとなる記述パターン（抽出条件）及びその領域に付すべきタグ（即ち、識別子としての要素名）を各要素毎に定義し

た「ＤＴＤ＋パターン情報（定義情報）」Ｒを、予め登録している。そして、「ＤＴＤ＋パターン情報」Ｒによって定義された各要素毎に、その要素の抽出条件に合致した領域を処理対象文書Ｔから抽出して、抽出した領域を「ＤＴＤ＋パターン情報」Ｒによって定義された各要素間の相互関係に従って組み合わせるとともに、「ＤＴＤ＋パターン情報」Ｒによって定義されているタグを各領域の前後に付す。このようにして、最終的には、各々タグが付された複数の領域からなる「構造化文書」Ｏを、生成して出力する。この「構造化文書」Ｏは、ＸＭＬ又はＳＧＭＬに従った構造を有しているので、一般のＸＭＬ／ＳＧＭＬシステムによって処理可能となる。

【００１２】

この「ＤＴＤ＋パターン情報」Ｒは、それ自体はテキスト形式で表されたファイルであるが、図７及び図１２に示されるように、上述した一般のＤＴＤと同様に、各要素の階層構造（即ち、一つの上位階層の要素内に複数の下位階層の要素が含まれてなる階層構造）を、ツリー形式で表すことができる。このようにツリー形式で表した場合に、最上位階層となる文書全体に対応した要素は「ルートノード」と称される。また、対象要素の直下の階層に存在する要素は当該対象要素にとっての「子ノード」と称され、逆に、対象要素の直上の階層に存在する要素は当該対象要素にとっての「親ノード」と称される。さらに、「子ノード」の「子ノード」は「孫ノード」と称される。また、同じ「親ノード」の「子ノード」の中では、ツリー図における上方に存在しているものが、下方に存在しているものにとっての「兄ノード」と称され、下方に存在しているものが、上方に存在しているものにとっての「弟ノード」と称される。特に、同じ「親ノード」の「子ノード」のうち最上方に存在しているものは、その「親ノード」にとっての「長子ノード」と称される。なお、同じ要素名の要素が（即ち、同じ構造の要素）が繰り返す場合には、その要素名には“＊”が付されることによって、『繰り返しあり（繰返構造）』の旨が指定される。

【００１３】

但し、この「ＤＴＤ＋パターン情報」Ｒは、各要素毎にその抽出条件として記述パターンが定義されている点において、一般のＤＴＤと異なる。この記述パタ

ーンの指定の仕方としては、抽出すべき領域の開始パターン及び終了パターンを文字列そのもの又は正規表現によって指定する仕方と、抽出すべき領域の全域を正規表現（そのルールの一部を図１１に示す）によって指定する仕方とが、利用可能である。前者の場合には、さらに、その開始パターン又は終了パターンを抽出すべき領域内に含むか、開始パターンの直後以降を抽出すべき領域とするか、終了パターンの直前までを抽出すべき領域とするかを、指定することが可能である。これらの様々な指定の仕方は、同一の「ＤＴＤ＋パターン情報」Ｒ内に混在していても良くなっている。なお、処理対象文書ＴがＢＮＦ（バッカス・ナウア・フォーム）による文法（そのルールの一部を図１４に示す）に従って記述されたソースプログラムリストである場合には、後述するように、ＢＮＦに従った「構文要素（シンタックス）」自体によって指定する仕方が、利用される。この場合には、「構文要素」の前後に存在するコメントをも合わせて抽出する旨を指定することも可能であり、さらに、このコメント部分についての子ノードに関しては、上述した文字列そのもの又は正規表現によって記述パターンを指定する仕方が、採用される。何れの場合においても、処理対象文書Ｔの全体を示す要素の抽出条件は、特殊な場合として、「文書（テキスト）全体」と指定される。以上のような記述パターンを様々に指定する「ＤＴＤ＋パターン情報」Ｒ内の情報を、以下、記述パターン情報という。

【００１４】

図７は、図８に示すような一般のテキストが処理対象文書Ｔである場合に適用される「ＤＴＤ＋パターン情報」Ｒの例を、ツリー構造で示した図である。この図７に示される例において、要素「ヘッダ」を抽出するための記述パターンは、“タイトル：”という文字列自体からなる記述パターンの直後から、『行頭から０個以上の空白の後“３．”という文字列があってその後に任意の文字が続く』記述パターンの直前までの領域が、抽出対象領域であることを示す。また、要素「ヘッダ」の子ノードである要素「日付」を抽出するための記述パターンは、要素「ヘッダ」の記述パターンによって抽出される領域中において、“対応日：”という文字列自体からなる記述パターンの直後から、その後最初の改行の直前までの領域が、抽出対象領域であることを示す。また、「繰り返しあり」の旨が指

定された要素「内容」を抽出するための記述パターンは、『行頭から０個以上の空白の後“４”乃至“９”の何れか及び“．”からなる文字列があってその後任意の文字が繰り返した後改行する』記述パターンの直後から、『行頭の直後に改行する』記述パターンの直前までの領域が、抽出対象領域であることを示す。

【００１５】

図７に示される「ＤＴＤ＋パターン情報」Ｒを参照して図８に示される処理対象文書Ｔに対する処理を行った場合に、抽出された領域が「ＤＴＤ＋パターン情報」Ｒによって定義された相互関係に従って階層化されてなるツリー構造を、図９に示す。このツリー構造において、上記要素「ヘッダ」として抽出された領域は“商談対応報告書～1997.02.17”であり、上記要素「日付」として抽出された領域は“1997.02.17”であり、上記要素「内容」として抽出された領域は、“以下のＳＤＡＳ～ＹＰＳ”及び“当社説明員による～回答も行う”の二つである。更に、図９に示されるツリー構造に基づいて、各要素に対応して抽出された領域の前後にその要素名をタグとして付すことによって作成された構造化文書Ｏを、図１０に示す。

【００１６】

図１２は、図１３に示されるようなソースプログラムリスト（より具体的には、Ｊａｖａソース）が処理対象文書Ｔである場合に適用される「ＤＴＤ＋パターン情報」Ｒの例を、ツリー構造で示した図である。なお、ソースプログラムリストが処理対象文書Ｔである場合には、構造文書化システム５は、この処置対象文書Ｔに含まれる各構文要素の範囲及び内容を、図１４にその一部が示されるＢＮＦ（バッカス・ナウア・フォーム）を定義した構文分解定義ファイルＢに従って、図１５に示すように解析する。そして、解析された構文要素がなす階層構造を、図１６に示されるようなツリー構造（構文木・コメント木）として、ＲＡＭ３上に構築する。図１４乃至図１６から明らかなように、ＢＮＦによると、例えば、ClassDefinitionには、Name（図１３，図１５の例では“顧客”）及び、MethodDefinition又はFieldDefinitionが含まれ、MethodDefinitionにも、Name（図１３，図１５の例では“信用ランク”）が含まれる。

【００１７】

図１２に示される「ＤＴＤ＋パターン情報」Ｒでは、要素「クラス仕様」を抽出するための記述パターン“Comment+ClassDefinition”は、ＢＮＦに定義された構文要素「ClassDefinition」に合致した領域，及び、この領域の直前に連続したコメントの領域が、抽出対象領域であることを示す。また、要素「クラス仕様」の子ノードである要素「作成者」を抽出するための記述パターンは、要素「クラス仕様」の記述パターンによって抽出されるコメント領域中において、“作成者”という文字列自体からなる記述パターンの直後から、その後最初の改行の直前までの領域が、抽出対象領域であることを示す。また、要素「クラス仕様」の子ノードである要素「クラス名」を抽出するための記述パターン“ClassDefinition.Name”は、要素「クラス仕様」の記述パターンによって抽出される構文要素領域中において、ＢＮＦに定義された構文要素「Name」に合致した領域が、抽出対象領域であることを示す。また、また、要素「クラス仕様」の子ノードである要素「メソッド仕様」を抽出するための記述パターン“Comment+MethodDefinition”は、要素「クラス仕様」の記述パターンによって抽出される構文要素領域中において、ＢＮＦに定義された構文要素「MethodDefinition」に合致した領域，及び、この領域の直前に連続したコメントの領域が、抽出対象領域であることを示す。また、要素「メソッド仕様」の子ノードである要素「メソッド名」を抽出するための記述パターン“MethodDefinition.Name”は、要素「メソッド仕様」の記述パターンによって抽出される構文要素領域中において、ＢＮＦに定義された構文要素「Name」に合致した領域が、抽出対象領域であることを示す。また、要素「クラス仕様」の子ノードである要素「説明」を抽出するための記述パターンは、要素「メソッド仕様」の記述パターンによって抽出されるコメント領域中において、“説明：”という文字列自体からなる記述パターンの直後から、その後の改行以外の任意の文字列からなる領域が、抽出対象領域であることを示す。また、要素「メソッド仕様」の子ノードである繰返構造を指定された要素「パラメタ」を抽出するための記述パターン“MethodDefinition.PrameterName”は、要素「メソッド仕様」の記述パターンによって抽出される構文要素領域中において、ＢＮＦに定義された構文要素「Prameter」に合致した全ての領域が、抽出対象領域であることを示す。

【００１８】

図１２に示される「ＤＴＤ＋パターン情報」Ｒを参照して図１３に示される処理対象文書Ｔに対する処理を行った場合に、抽出された領域が「ＤＴＤ＋パターン情報」Ｒによって定義された相互関係に沿って階層化されてなるツリー構造を、図１７に示す。このツリー構造において、上記要素「クラス仕様」として抽出される領域は

```
“/** COPYRIGHT Fujitsu LTD
* 作成者　藤川　泰之（富士通）
* 更新者　原田　義之（富士通）
* 更新者　和田　憲明（富士通）
*/
public class 顧客 {
/*
 *説明：資本金から信用度を割り出す。
*/
        public String 信用ランク(
                int 現在借入金
                long 公定歩合)
        {
                :
                :
        }
//説明:資本金。
        【００１９】
        public static int 資本金:
} ”
```

である。

【００２０】

また、上記要素「作成者」として抽出される領域は“* 作成者　藤川　泰之（

富士通）”であり、上記要素「クラス名」として抽出される領域は“顧客”であり、上記要素「メソッド仕様」として抽出される領域は

“/*

*説明：資本金から信用度を割り出す。

*/

```
public String 信用ランク(
        int 現在借入金
        long 公定歩合)
 {
        :
        :
} ”
```

である。

【００２１】

また、上記要素「メソッド名」として抽出される領域は“信用ランク”であり、上記要素「説明」として抽出される領域は、“資本金から信用度を割り出す。”であり、上記要素「パラメタ」として抽出される領域は、“int 現在借入金”及び“long 公定歩合”の二つである。

【００２２】

図１に戻り、以上のようにして構造文書化システム５によって参照される「ＤＴＤ＋パターン情報」Ｒは、ＤＴＤ＋パターン編集システム６によって編集される。このＤＴＤ＋パターン編集システム６は、図１８に示すようなＧＵＩ（編集画面）を有するテキストエディタである。このＤＴＤ＋パターン編集システム６の編集画面の左半分はＤＴＤツリー構造リストボックス６１となっており、その右半分は項目入力部６２となっている。また、この画面の下縁近傍には、削除ボタン６３，キャンセルボタン６４，更新終了ボタン６５，内容反映ボタン６６，子として追加ボタン６７，及び、弟として追加ボタン６８が、並べて表示されている。

【００２３】

ＤＴＤツリー構造リストボックス６１は、編集中の「ＤＴＤ＋パターン情報」Ｒによって定義される各要素の要素名を、各要素間の階層構造を示すツリー図として表示するリストボックスである。このＤＴＤツリー構造リストボックス６１に表示されている何れかの要素名が入力装置（マウス）８を介してオペレータによってクリックされると、そのクリックされた要素名が示す要素が処理対象として選択され、その表示色が変更されるとともに（図１８の例においては要素名「タイトル」の表示色が変更されている）、その要素名が示す要素についての現在の設定内容が、項目入力部６２の各テキストボックスやチェックボックスやオプションボタンに表示される。

【００２４】

この項目入力部６２は、要素名テキストボックス６２１，繰り返しありチェックボックス６２１０，パターンの意味オプションボタン６２２，前後空白除去チェックボックス６２２０，パターン／開始パターン指定部６２４，終了パターン指定部６２５，親に対する範囲制限オプションボタン６２６，出力タグ名テキストボックス６２７を含む。

【００２５】

要素名テキストボックス６２１は、現在選択されている要素の要素名を表示する（記述される）テキストボックスである。また、繰り返しありチェックボックス６２１８は、現在選択されている要素に繰り返しあり（繰返構造）とするか否かを表示する（指定される）するチェックボックスである。また、パターンの意味オプションボタン６２２は、現在選択されている要素における記述パターンの指定の仕方が、要素の開始パターン及び終了パターンを指定する仕方であるか要素全体の記述パターンそのものを指定する仕方であるかを表示する（指定される）オプションボタンである。また、前後空白除去チェックボックス６２２０は、現在選択されている要素を抽出する際に抽出対象領域の前後に空白が含まれていたならばこれを除去するか否かを表示する（指定される）チェックボックスである。行頭文字除去テキストボックス６２３は、現在選択されている要素を抽出する際に抽出対象領域の行頭に含まれていたならば削除される文字列を表示する（記述される）テキストボックスである。

【００２６】

パターン／開始パターン指定部６２４は、現在選択されている要素全体の記述パターン（パターンの意味オプションボタン６２２においてパターンそのものが指定されている場合）又は開始パターン（パターンの意味オプションボタン６２２において開始と終了が指定されている場合）の内容を表示する（指定される）領域である。このパターン／開始パターン指定部６２４は、パターンタイプオプションボタン６２４１，コメント処理チェックボックス領域６２４２，内容にパターンを含むチェックボックス６２４３，構文要素名参照ボタン６２４４，パターン記述テキストボックス６２４５を含む。

【００２７】

パターンタイプオプションボタン６２４１は、対象記述パターンが文字列そのものであるか正規表現であるか構文要素名があるかを表示する（指定される）オプションボタンである。また、コメント処理チェックボックス領域６２４２は、パターンタイプオプションボタン６２４１にて構文要素名が選択されている場合において構文要素の前方に連続するコメントを抽出するか否かを表示する（指定される）チェックボックス，及び、同様に構文要素の後方に連続するコメントを抽出するか否かを表示する（指定される）チェックボックスを含む領域である。内容にパターンを含むチェックボックス６２４３は、パターンの意味オプションボタン６２２にて開始と終了が選択されている場合において、記述パターンに対応した文字列を抽出対象領域に含むか否かを表示する（指定される）チェックボックスである。また、構文要素名参照ボタン６２４４は、パターンタイプオプションボタン６２４１にて構文要素名が選択されている場合において、構文分解定義ファイルＢに定義されている各構文要素名及びその定義内容のリストを表示するためにクリックされるボタンである。また、パターン記述テキストボックス６２４５は、現在選択されている要素における全体の記述パターン（パターンの意味オプションボタン６２２においてパターンそのものが指定されている場合）又は開始パターン（パターンの意味オプションボタン６２２において開始と終了が指定されている場合）自体を表示する（記述される）テキストボックスである。

【００２８】

終了パターン指定部６２５は、現在選択されている要素における終了パターン（パターンの意味オプションボタン６２２において開始と終了が指定されている場合）の内容を表示する（指定される）領域である。この終了パターン指定部６２５は、パターンタイプオプションボタン６２５１，内容にパターンを含むチェックボックス６２５５，構文要素名参照ボタン６２５４，パターン記述テキストボックス６２５５を含む。これらの機能は、パターン／開始パターン指定部６２４のものと全く同じであるので、その説明を省略する。

【００２９】

親に対する範囲制限オプションボタン６２６は、現在選択されている要素の親ノードに指定された記述パターンが構文要素である場合、当該要素の検索範囲が親ノードに対応する領域の全体であるか（無し）、親ノードに対応する領域における構文要素の部分であるか（構文要素）、親ノードに対応する領域における構文要素の前方に連続したコメント部であるか（前方コメント）親ノードに対応する領域における構文要素の後方に連続したコメント部であるか（後方コメント）を表示する（指定される）オプションボタンである。

【００３０】

出力タグ名テキストボックス６２７は、現在選択されている要素に対応する領域が抽出された後にその領域の前後に付加されるタグ（通常は要素名テキストボックス６２１に表示された要素名と同じ）を表示する（記述される）テキストボックスである。

【００３１】

何れかの要素が選択された状態において、オペレータが削除ボタン６３をクリックすると、その要素の設定内容（ＤＴＤ構造及び記述パターン情報）が消去される（この場合には、項目入力部６２内の各テキストボックス，チェックボックス及びオプションボタンは、全て空欄となる。）。

【００３２】

また、何れかの要素が選択された状態において、オペレータがキャンセルボタン６４をクリックすると、その要素の選択が解除される（この場合には、項目入力部６２内の各テキストボックス及びオプションボタンが全て空欄となるとも

に、ＤＴＤツリー構造リストボックス６１内における当該要素の要素名の表示色が元に戻る。）。

【００３３】

また、何れかの要素が選択された状態において、オペレータが項目入力部６２内の何れかのテキストボックス内の記述又は何れかのチェックボックス又はオプションボタンにおけるチェック内容を変更した後に、内容反映ボタン６６をクリックすると、その要素の設定内容が、項目入力部６２に現在表示されている内容に変更される。

【００３４】

また、何れかの要素が選択された状態において、オペレータが項目入力部６３内における少なくとも要素名テキストボックス６２１の記述を変更した後に、子として追加ボタン６７をクリックすると、その要素の子ノードとして、項目入力部６２に現在表示されている設定内容を有する要素が、追加される。

【００３５】

また、何れかの要素が選択された状態において、オペレータが項目入力部６３内における少なくとも要素名テキストボックス６２１の記述を変更した後に、弟として追加ボタン６８をクリックすると、その要素の弟ノードとして、項目入力部６２に現在表示されている設定内容を有する要素が、追加される。

【００３６】

さらに、ＤＴＤツリー構造リストボックス６１内に表示されている要素名が入力装置８を介してオペレータによってドラッグされて、他の何れかの要素名上にドロップされると、ドラッグされた要素名が示す要素の設定内容が、ドロップ先の要素名が示す要素の子ノードとなるように、変更される。

【００３７】

最後に、オペレータが内容反映ボタン６６をクリックすると、現在における各要素の設定内容に従って、ＤＴＤパターン情報Ｒが作成又は更新される。

【００３８】

以上のような編集画面及びこの編集画面に関連付けられた機能を有するＤＴＤ＋パターン編集システム６を用いて、オペレータは、自在にＤＴＤ＋パターン情

報Ｒを編集することができる。オペレータは、このＤＴＤ＋パターン編集システム６を用いて、ゼロからＤＴＤ＋パターン情報Ｒを作成することもできるが、図１に示したＤＴＤ＋パターン作成支援システム７によって作成されたＤＴＤ＋パターン情報Ｒを、ＤＴＤ＋パターン編集システム６を用いて編集することによって完成させても良い。

【００３９】

このＤＴＤパターン作成支援システム７は、図２４に示すようなＧＵＩ（選択画面）を有するテキストエディタである。このＤＴＤパターン作成支援システム７は、図２５及び図２６に示すような複数の典型パターン定義情報Ｓを有している。各典型パターン定義情報Ｓは、定型的な文書に頻繁に表れる典型的な文字列パターン（以下、「典型パターン」という）を要素として抽出するための記述パターン情報の雛形を、定義している情報である。即ち、図２５，図２６に示すように、各典型パターン定義情報Ｓは、その典型パターンの概略構造を特定する構造特定部Ｓ１，構造特定部Ｓ１中において典型パターンの概略構造を構成する個々のエレメント（隅付き括弧で囲まれた部分）として用いられ得る文字の種類を正規表現で特定する文字種特定部Ｓ２，及び、ＤＴＤパターン情報Ｒにおける各要素毎の記述パターン情報の雛形である雛形部Ｓ３から、構成されている。

【００４０】

図２５の例は、例えば、

“会社名：富士通株式会社”

の様に、『行頭から０個以上の空白の後に、項目名（タイトル），区切り文字（区切り），具体的内容（内容）と続き、行末に到る』という典型パターンを一つの要素として抽出するための記述パターン情報の雛形を定義しているので、構造特定部Ｓ１には、“《行頭》＊［タイトルパターン（項目名に対応）］＊［区切りパターン（区切り文字に対応）］＊［内容パターン（具体的内容に対応）］＊《行末》”と、特定されている。また、文字種特定部Ｓ２では、［タイトルパターン］及び［内容パターン］については“《改行以外》＋”と特定され、［区切りパターン］については“；：／（）”と特定されている。また、雛形部Ｓ３では、パターン指定方式が“開始と終了”と特定され、開始パターンが正規表現に

よる“《行頭》＊［タイトル文字列１］｜［タイトル文字列２］＊［区切り文字列１］｜［区切り文字列２］＊”と特定され、終了パターンが正規表現による“＊《行末》”と特定されている。［タイトル文字列１］及び［タイトル文字列２］は、項目名として採り得る記述候補が代入される部分である。同様に、［区切り文字列１］及び［区切り文字列２］は、区切り文字列として採り得る記述候補が代入される部分である。

【００４１】

図２６の例は、一つの親ノード及び複数の子ノードとして抽出される典型パターンに用いられる典型パターン定義情報Ｓであるので、雛形部Ｓ３としては、親ノードを抽出するための記述パターン情報の雛形（以下、「親ノード雛形Ｓ３ａ」という）と、構造特定部Ｓ１に記載された各［タイトルパターン］に夫々対応した子ノードを夫々抽出するための複数個の記述パターン情報の雛形（以下、「子ノード雛形Ｓ３ｂ」という）とが、備えられている。従って、親ノード雛形Ｓ３ａには、各子ノードの要素名が夫々代入される［タイトルパターン１］～［タイトルパターン５］が、含まれている。また、各子ノード雛形Ｓ３ｂでは、兄ノードとの関係が、“順序性＝有り”と特定されている。

【００４２】

図２４に示す選択画面は、ルート文書要素名テキストボックス７１，サンプルプレビューボックス７２，ツリーリストボックス７３，及び、典型パターン選択領域７４を、含んでいる。この典型パターン選択領域７４は、各典型パターン定義情報Ｓに対応付けられた複数の典型パターン選択ボタン７４１を含んでいる。そして、各典型パターン選択ボタン７４１の表面には、それに対応付けられた典型パターン定義情報Ｓにおける構造特定部Ｓ１の特定内容を端的に示す文字列が、表示されている。例えば、図２５に示される典型パターン定義情報Ｓは、図２４における一番上の典型パターン選択ボタン７４１に対応付けられているので、この典型パターン選択ボタン７４１には、文字列“タイトル：ＮＮＮＮＮＮＮＮＮ”と、表示されている。

【００４３】

ＤＴＤ＋パターン作成支援システム７は、処理対象文書Ｔのサンプルを読み込

むと、そのテキスト内容を、サンプル表示ダイアログ７２内に表示する。ルート文字要素名テキストボックス７１は、サンプル表示ダイアログ７２内に表示されたサンプルに基づいて作成されるＤＴＤ＋パターン情報Ｒにおけるルートノードの要素名が記述されるテキストボックスである。このルート文字要素名テキストボックス７１内にルートノードの要素名がオペレータによって書き込まれることにより、ＤＴＤ＋パターン作成支援システム７は、このルートノードの要素名のみを原始内容とするＤＴＤ＋パターン情報Ｒを生成する。

【００４４】

ＤＴＤ＋パターン作成支援システム７は、サンプル表示ダイアログ７２内に表示されたテキストにおける何れかの行がドラッグされて選択された後に、何れかの典型パターン選択ボタン７４１がクリックされると、その典型パターン選択ボタン７４１に対応付けられた典型パターン定義情報Ｓを読み出し、その構造特定部Ｓ１内で特定された典型パターンの概略構造を選択された行に当てはめて、その概略構造を構成する各エレメントに相当する文字列を抽出する。そして、抽出した各エレメントに関する文字列の文字種を、文字種特定部Ｓ２にて特定された文字種に変換する。そして、変換後における各エレメントに関する文字列を、雛形部Ｓ３の記述パターン情報の雛形における各［］に、代入する。このようにして、ＤＴＤ＋パターン作成支援システム７は、ルート文書要素名テキストボックス７１に記述された要素名を有するルートノードの子ノード（若しくは、子ノード及び孫ノード）を抽出するための記述パターン情報を作成し、この指定内容をＤＴＤ＋パターン情報Ｒに追加する。

【００４５】

ツリーリストボックス７３は、作成中のＤＴＤ＋パターン情報Ｒに含まれる各要素の要素名がその階層構造を示すツリー図として表示されるリストボックスである。従って、オペレータが、サンプル表示ダイアログ７２内に表示されているテキスト中における何れか行をドラッグして何れかの典型パターン選択ボタン７４１をクリックする毎に、ツリーリストボックス７３中に表示されているルートノードの下位階層に、子ノード（若しくは、子ノード及び孫ノード）の要素名が追加される。

（構造文書化システムの詳細構成及び処理内容）

次に、上述した構造文書化システム５の詳細な構成を、その処理内容に沿って説明する。図３は、この構造文書化システム５の詳細な構成（構造文書化システム５を構築するプログラムの詳細なモジュール構成）を示すブロック図である。また、図４乃至図６は、この構造文書化システム５の処理内容（構造文書化システム５を構築するプログラムによるＣＰＵ１の処理内容）を示すフローチャートである。

【００４６】

図３に示されるように、構造文書化システム５は、ＤＴＤ＋パターンツリー作成部５１，全体コントロール部５２，パターン検索部５３，及び、構文木分解部５４から、構成されている。更に、パターン検索部５３は、文字列検索部５３１，正規表現検索部５３２及び構文要素検索部５３３を、含んでいる。

【００４７】

構文木分解部５４は、処理対象文書Ｔがソースプログラムリスト（但し、ＢＮＦによる文法に従って記述されていることが条件）である場合に起動し、この処理対象文書Ｔの内容を構文分解定義ファイルＢに従って解析して、解析した処理対象文書Ｔの構文構造に応じて、図１６に示したような構文木・コメント木５７をＲＡＭ３上に構築する。

【００４８】

一方、ＤＴＤ＋パターンツリー作成部５１は、オペレータによって選択されたＤＴＤ＋パターン情報Ｒを読み込んで（読込部に相当）、その内容を解析することによって、図７及び図１２に示したようなＤＴＤ＆パターンツリー５５を、ＲＡＭ３上に構築する。

【００４９】

全体コントロール部５２は、ＤＴＤ＋パターンツリー作成部５１が作成したＤＴＤ＋パターンツリー５５における各要素のパターン記述情報を順次読み出して、読み出したパターン記述情報に沿った処理対象文書Ｔ中の領域の抽出を、パターン検索部５３に対して依頼する。この際に、全体コントロール部５２は、その要素について繰り返し有りと指定されている場合には、パターン検索部５３が検

索結果を報告できなくなるまで、その要素のパターン記述情報に沿った領域の抽出を、パターン検索部５３に対して依頼する。そして、全体コントロール部５２は、各要素毎になした抽出依頼に応じてパターン検索部５３が抽出して来た処理対象文書Ｔ中の領域を、ＤＴＤ＋パターンツリー５５における各要素の位置（即ち、「ＤＴＤ＋パターン情報」Ｒ中のＤＴＤ）に従って、図９及び図１７に示したような出力結果ツリー５６として組み上げ、最後に、出力結果ツリー５６における各要素に対応した領域の前後にその要素に対応したタグを付加することによって、図１０に示したような構造化文書Ｏを出力する（構造化文書生成部に相当）。

【００５０】

パターン検索部５３は、全体コントロール部５２から抽出を依頼された要素の記述パターン情報における記述パターン（要素全体のパターン，若しくは、開始パターン及び終了パターン）の種類に対応したパターン検索部５３（パターン記述が文字列そのものである場合には文字列検索部５３１，記述パターンが正規表現である場合には正規表現検索部５３２，記述パターンが構文要素である場合には構文要素検索部５３３）を呼び出して、その記述パターンに対応した文字列の検索を命じる。この際に、パターン検索部５３は、抽出対象の要素の親ノードについて既に抽出した領域を、検索対象範囲として指定する。また、抽出対象の要素に順序性有りと指定されている場合には、親ノードについて既に抽出した領域における兄ノードについて既に抽出した領域の後の部分を、検索対象範囲として指定する。また、抽出対象の要素に繰り返し有りの指定がなされている場合において、全体コントロール部５２から同じ要素の抽出を２回目以降に依頼された場合には、親ノードについて既に抽出した領域におけるその要素について前回に抽出した領域の後の部分を、検索対象範囲として指定する。なお、パターン検索部５３は、開始パターンと終了パターンとで記述パターンの種類が異なるのであれば、夫々の記述パターンに対応して文字列検索部５３１及び正規表現検索部５３２を呼び出して、夫々の記述パターンに対応した文字列の検索を命じる。

【００５１】

パターン検索部５３は、文字列検索部５３１，正規表現検索部５３２又は構文

要素検索部５３３が検索結果を報告して来ると（文字列検索部５３１及び正規表現検索部５３２に開始パターンと終了パターンとに対応した文字列の検索を命じた場合ににおいては、両検索部５３１，５３２からの検索結果の報告が揃うと）、その検索結果を参照して、その要素に対応する領域（要素全体の記述パターンが指定されている場合には検索された文字列，開始パターン及び終了パターンが指定されている場合には、検索された文字列に挟まれた領域，但し、開始パターン又は終了パターンについて「内容にパターンを含む」と指定されている場合にはその記述パターンをも含む領域）を処理対象文書Ｔから抽出して、抽出した領域を全体コントロール部５２に通知する（検索部に相当）。

【００５２】

文字列検索部５３１は、パターン検索部５３から命じられた記述パターンと全く同じ文字列を検索し、正規表現検索部５３２は、パターン検索部５３から命じられた記述パターンが示す正規表現に合致する文字列を検索し、構文要素検索部は、パターン検索部５３から命じられた記述パターンと同じ構文要素（又は／及び、その前又は後に連続したコメント部）を検索し、パターン検索部５３に報告する。

【００５３】

以上のような各モジュールから構成される構造文書化システム５は、入力装置８を介してオペレータによって入力される開始コマンドによって起動し、同じくオペレータによる入力によって処理対象文書Ｔ及びＤＴＤ＋パターン情報Ｒが選択されることによって、図４に示す手順にて処理を開始する。

【００５４】

図４において、スタート後最初のＳ００１では、ＤＴＤ＋パターンツリー作成部５１が、選択されたＤＴＤ＋パターン情報Ｒを、ハードディスク２からＲＡＭ３上に読み込む。

【００５５】

次のＳ００２では、ＤＴＤ＋パターンツリー作成部５１が、Ｓ００１にて読み込んだＤＴＤ＋パターン情報Ｒに基づいて、ＤＴＤ＋パターンツリー５５を、ＲＡＭ３上で構築する。

【００５６】

次のＳ００３では、全体コントロール部５２が、選択された処理対象文書Ｔをハードディスク２からＲＡＭ３上に読み込む。

【００５７】

次のＳ００４では、全体コントロール部５２が、Ｓ００２にて作成されたＤＴＤ＋パターンツリー５５が、構文要素からなる記述パターンを含むか否かをチェックする。そして、ＤＴＤ＋パターンツリー５５が構文要素からなる記述パターンを含まない場合には、全体コントロール部５２は、Ｓ００６において処理対象文書Ｔ自体を検索対象として決定した後に、処理をＳ００７に進める。これに対して、ＤＴＤ＋パターンツリー５５が構文要素からなる記述パターンを含む場合には、全体コントロール部５２は、Ｓ００５において、構文分解定義ファイル８を読み込んで、この構文分解定義ファイルＢを参照することによって、処理対象文書Ｔに基づいて構文木・コメント木５７を作成し、この構文木・コメント木５７を検索対象として決定した後に、処理をＳ００７に進める。

【００５８】

Ｓ００７では、全体コントロール部５２が、ＤＴＤ＋パターンツリー５５に従って出力結果ツリー５６を作成する処理を、実行する。図５及び図６は、このＳ００７にて実行される出力結果ツリー作成処理サブルーチンを示すフローチャートである。このサブルーチンに入って最初のＳ１０１では、全体コントロール部５２が、ＤＴＤ＋パターンツリー５５におけるルートノードに対応した領域が処理対象文書全体であると決定して、処理対象文書Ｔ全体をルートノードに対応した抽出結果とする出力結果ツリー５６を生成する。

【００５９】

次のＳ１０２では、全体コントロール部５２は、ＤＴＤ＋パターンツリー５５におけるルートノードの長子ノードを、処理対象ノードとする。次に、全体コントロール部５２は、Ｓ１０３乃至Ｓ１１３のループ処理を実行する。このループ処理に入って最初のＳ１０３では、全体コントロール部５２は、ＤＴＤ＋パターンツリー５５における処理対象ノードから、その要素に指定された記述パターン情報を取り出す。

【００６０】

次のＳ１０４では、全体コントロール部５２は、処理対象ノードの親ノードに対応した領域内（構文木・コメント木５７にあっては親ノードの下位階層）を、Ｓ１０３にて取り出した記述パターンに合致する領域（要素全体の記述パターンが指定されている場合には検索された文字列，開始パターン及び終了パターンが指定されている場合には、検索された文字列に挟まれた領域，但し、開始パターン又は終了パターンについて「内容にパターンを含む」と指定されている場合にはその記述パターンをも含む領域）の検索対象範囲として、決定する。

【００６１】

次のＳ１０５では、パターン検索部５３は、親ノードの領域内における検索開始位置を、処理対象ノードの特性（順序性の有無，兄ノードの有無，繰り返し指定があるノードに対する２回目以降の処理であるか否か）に従って、決定する。即ち、順序性が有り且つ兄ノードがある場合（但し、繰り返し指定があるが当該処理対象ノードに対する２回目以降の処理である場合を除く）には、パターン検索部５３は、Ｓ１０６において、直前の兄ノードに対応した検索済み領域の後から検索することとする。また、繰り返し指定があるが当該処理対象ノードに対する２回目以降の処理である場合には、パターン検索部５３は、Ｓ１０７において、当該処理対象ノードに関して前回の処理で検索した領域の後ろから検索することとする。また、繰り返し指定無し且つ順序性無しの場合や、その他の場合には、パターン検索部５３は、Ｓ１０８において、親ノードの先頭から検索することとする。

【００６２】

何れの場合においても、次のＳ１０９において、パターン検索部５３は、処理対象ノードの記述パターン情報における記述パターンの指定方法（要素全体の記述パターンを指定するか要素の開始パターン及び終了パターンを指定するか）及び表現方法（文字列のものを指定するか文字列の正規表現を指定するか）に従って、検索対象領域から、指定された記述パターンに合致した領域を検索する。この検索によって抽出された結果は、全体コントロール部５２に通知される。

【００６３】

次のＳ１１０では、全体コントロール部５２は、Ｓ１０９での検索の結果として処理対象ノードの記述パターンに合致した領域が検索対象領域から抽出されたか否かをチェックする。そして、合致した領域が抽出されがならば、全体コントロール部５２は、Ｓ１１１において、抽出された領域の文字列を内容とするノードを、出力結果ツリー５６における親ノードの下位階層に追加する。

【００６４】

次のＳ１１２では、全体コントロール部５２は、処理対象ノードに子ノードがあるか否かをチェックする。そして、処理対象ノードに子ノードがあるならば、全体コントロール部５２は、Ｓ１１３において、現在の処理対象ノードの長子ノードを、新たな処理対象ノードとして、処理をＳ１０３に戻す。

【００６５】

以上説明したＳ１０３乃至Ｓ１１３のループ処理を繰り返した結果、Ｓ１０９での検索の結果として処理対象ノードの記述パターンに合致した領域が検索対象領域から抽出されなかったとＳ１１０にて判定した場合には、全体コントロール部５２は、Ｓ１１４において、現在の処理対象ノードに対応した領域は無いものとして、空文字列を内容とするノードを出力結果ツリー５６における親ノードの下位階層に追加する。このＳ１１４の完了後、全体コントロール部５２は、処理をＳ１１６に進める。

【００６６】

また、上述したＳ１０３乃至Ｓ１１３のループ処理を繰り返した結果、処理対象ノードに子ノードが無いとＳ１１２にて判定した場合（処理対象ノードがいわゆる葉ノードである場合）には、全体コントロール部５２は、処理をＳ１１５に進める。

【００６７】

Ｓ１１５では、全体コントロール部５２は、処理対象ノードに繰り返し指定があるか否かをチェックする。そして、繰り返し指定がある場合には、全体コントロール部５２は、処理対象ノードをそのままにして、処理をＳ１０３に戻す。

【００６８】

これに対して、処理対象ノードに繰り返し指定がないと判定した場合には、全

体コントロール部５２は、処理をＳ１１６に進める。

【００６９】

Ｓ１１６では、全体コントロール部５２は、処理対象ノードに弟ノードがあるか否かをチェックする。そして、弟ノードがある場合には、全体コントロール部５２は、Ｓ１１７において、次弟ノードを新たな処理対象ノードとして、処理をＳ１０３に戻す。

【００７０】

これに対して、処理対象ノードに弟ノードが無いと判定した場合には、全体コントロール部５２は、Ｓ１１８にて現在の処理対象ノードの親ノードを仮の処理対象ノードとして、処理をＳ１１９に進める。Ｓ１１９では、全体コントロール部５２は、仮の処理対象ノードがルートノードであるか否かをチェックする。そして、仮の処理対象ノードがルートノードではないならば、全体コントロール部５２は、処理をＳ１１５に戻す。この場合、全体コントロール部５２は、仮の処理対処ノードに繰り返し指定があるか否かをチェックして、繰り返し指定があれば当該仮の処理対象ノードを本来の処理対象ノードとして扱って、処理をＳ１０３に戻す。これに対して、仮の処理対象ノードに繰り返し指定がなければ、当該仮の処理対象ノードが弟ノードを有している否かをチェックする。そして、当該仮の処理対象ノードが弟ノードを有していれば、当該弟ノードを新たな処理対象ノードとし（Ｓ１１７）、弟ノードを有していなければ、当該仮の処理対象ノードの更に親ノードを仮の処理対象ノードとする（Ｓ１１８）。

【００７１】

以上説明したＳ１０３乃至Ｓ１１９の処理を繰り返すことによって、ＤＴＤ＋パターンツリー５５を構成する全てのノードに基づいた検索が行われる。そして、全ノードに基づいた検索が完了すると、Ｓ１１９にて仮の処理対象ノードがルートノードであるとの判定がなされ、この出力結果ツリー作成サブルーチンが終了し、処理を図４のメインルーチンに戻る。従って、この時点において、出力結果ツリー５６が完成する。

【００７２】

処理が戻された図４のメインルーチンでは、処理は、Ｓ００７からＳ００８に

進む。このＳ００８では、全体コントロール部５２は、Ｓ００７での処理の結果完成した出力結果ツリー５６に基づいて、構造化文書Ｏを作成する。具体的には、全体コントロール部５２は、子ノードを有していないノード（いわゆる葉ノード）に対応した領域の前後に、そのノード（要素）に対応付けられたタグを付加する。次に、兄弟ノード同士をひとまとめにして、その全体の前後に、これらノードに共通した親ノードに対応付けられたタグを付加する。このようにして、最下位階層ノードから上位ノードへと順次タグを付加して行き、最終的に、ルートノードに対応付けられたタグを付加することによって、構造化文書Ｏが完成する。全体コントロール部５２は、このようにして完成した構造化文書Ｏを、ハードディスク２及びディスプレイ４に出力する。

【００７３】

次のＳ００９では、全体コントロール部５２は、Ｓ００１にて読み込んだＤＴＤ＋パターン情報Ｒに基づいて処理すべき他の処理対象文書Ｔがオペレータによって選択されているか否かをチェックする。そして、他の処理対象文書Ｔがオペレータによって選択されていない場合には、全体コントロール部５２は、処理をＳ００３に戻す。

【００７４】

これに対して、他の処理対象文書Ｔがオペレータによって選択されている場合には、全体コントロール部５２は、Ｓ０１０において、現在参照しているＤＴＤ＋パターン情報Ｒを変更する旨がオペレータによって入力されているか否かをチェックする。そして、ＤＴＤ＋パターン情報Ｒを変更する旨が入力されている場合には、全体コントロール部５２は、処理をＳ００１に戻す。これに対して、ＤＴＤ＋パターン情報Ｒを変更する旨が入力されていない場合には、構造文書化システム５による処理が終了する。

（構造文書化システムの動作例）

次に、以上の様な手順で処理を実行する構造文書化システム５による具体的な動作例を、説明する。

【００７５】

いま、図１９に示すような内容を有するＤＴＤ＋パターン情報Ｒがオペレータ

によって選択されるとともに、図２１に示すような内容を有する処理対象文書Ｔがオペレータによって選択されたとする。すると、構造文書化システム５のＤＴＤ＋パターンツリー作成部５１は、このＤＴＤ＋パターン情報Ｒの内容を解析して、図２０に示すようなＤＴＤ＋パターンツリー５５を作成する（Ｓ００１，Ｓ００２）。

【００７６】

全体コントロール部５２は、このＤＴＤ＋パターンツリー５５を参照し、最初に、ルートノード「開発履歴」に対応する領域が、この処理対象文書Ｔの全体であると決定する（Ｓ１０１）。次に、全体コントロール部５２は、ルートノードの子ノートを、順を追って処理対象ノードとしていく（Ｓ１０２，Ｓ１０３～Ｓ１１３）。

【００７７】

先ず、全体コントロール部５２は、ルートノードの長子ノード「初版情報」を、処理対象ノードとする（Ｓ１０２）。そして、全体コントロール部５２は、ＤＴＤ＋パターンツリー５５から、このノード「初版情報」についての記述パターン情報を参照して（Ｓ１０３）、親ノード「開発履歴」に対応する領域（処理対象文書Ｔの全体）を検索対象とする（Ｓ１０４）。そして、この記述パターン情報には繰り返し及び順序性が共に無しと指定されているので、パターン検索部５３は、親ノード「開発履歴」に対応する領域の先頭から、検索を開始する（Ｓ１０８，Ｓ１０９）。この検索において、この記述パターン情報にはパターン指定方法が開始及び終了と指定され、開始パターンが文字列そのものによる“初版作成者”、終了パターンが正規表現による“《行末》”と指定されているので、

“藤川　泰之　：　1999.01.01”

の部分が、記述パターン情報に合致した領域として検出される。従って、この領域が、ノード「初版情報」に対応した領域として抽出され、出力結果ツリー５６に追加される（Ｓ１１１）。

【００７８】

次に、全体コントロール部５２は、その長子ノード「作成者」を、新たな処理対象ノードとする（Ｓ１１２，Ｓ１１３）。そして、全体コントロール部５２は

、ＤＴＤ＋パターンツリー５５から、このノード「作成者」についての記述パターン情報を参照して（Ｓ１０３）、親ノード「初版情報」に対応する領域

“藤川　泰之　：　1999.01.01”

を検索対象とする（Ｓ１０４）。そして、この記述パターン情報には繰り返し及び順序性が共に無しとして指定されているので、パターン検索部５３は、親ノード「初版情報」に対応する領域の先頭から、検索を開始する（Ｓ１０８，Ｓ１０９）。この検索において、この記述パターン情報にはパターン指定方法が開始及び終了と指定され、開始パターンが正規表現による“《領域の先頭》”、終了パターンが文字列そのものによる“：”と指定されているので、

“藤川　泰之”

の部分が、記述パターン情報に合致した領域として検出される。従って、この領域が、ノード「作成者」に対応した領域として抽出され、出力結果ツリー５６に追加される（Ｓ１１１）。

【００７９】

このノード「作成者」には子ノードがなく（Ｓ１１２）、その記述パターン情報には繰り返しが無しと指定されているので（Ｓ１１５）、全体コントロール部５２は、その次弟ノード「作成日時」を、新たな処理対象ノードとする（Ｓ１１６，Ｓ１１７）。そして、全体コントロール部５２は、ＤＴＤ＋パターンツリー５５から、このノード「作成日時」についての記述パターン情報を参照して（Ｓ１０３）、親ノード「初版情報」に対応する領域

“藤川　泰之　：　1999.01.01”

を検索対象とする（Ｓ１０４）。そして、この記述パターン情報には繰り返しが無しと指定されているが順序性が有りと指定されているので、パターン検索部５３は、兄ノード「作成者」の後から、検索を開始する（Ｓ１０６，Ｓ１０９）。この検索において、この記述パターン情報にはパターン指定方法が開始及び終了と指定され、開始パターンが文字列そのものによる“：”、終了パターンが正規表現による“《行末》”、と指定されているので、

“1999.01.01”

の部分が、記述パターン情報に合致した領域として検出される。従って、この領

域が、ノード「作成日時」に対応した領域として抽出され、出力結果ツリー５６に追加される（Ｓ１１１）。

【００８０】

このノード「作成者」には子ノードがなく（Ｓ１１２）、その記述パターン情報には繰り返しが無しと指定されており（Ｓ１１５）、しかも弟ノードがないので（Ｓ１１６）、全体コントロール部５２は、親ノード「初版情報」の次弟ノード「更新履歴」を、新たな処理対象ノードとする（Ｓ１１８，Ｓ１１９，Ｓ１１５～Ｓ１１７）。そして、全体コントロール部５２は、ＤＴＤ＋パターンツリー５５から、このノード「更新履歴」についての記述パターン情報を参照して（Ｓ１０３）、親ノード「開発履歴」に対応する領域（処理対象文書Ｔの全体）を検索対象とする（Ｓ１０４）。そして、この記述パターン情報には順序性有りと指定されているので、パターン検索部５３は、兄ノード「初版情報」の後から、検索を開始する（Ｓ１０６，Ｓ１０９）。この検索において、この記述パターン情報にはパターン指定方法が開始及び終了と指定され、開始パターンが文字列そのものによる“更新履歴”、終了パターンが正規表現による“《行末》”、と指定されているので、

“1999.12.16　／　第1.1版”

の部分が、記述パターン情報に合致した領域として最初に検出される。従って、この領域が、ノード「作成日時」に対応した領域として抽出され、出力結果ツリー５６に追加される（Ｓ１１１）。

【００８１】

次に、全体コントロール部５２は、その長子ノード「更新日付」を、新たな処理対象ノードとする（Ｓ１１２，Ｓ１１３）。そして、全体コントロール部５２は、ＤＴＤ＋パターンツリー５５から、このノード「更新日付」についての記述パターン情報を参照して（Ｓ１０３）、親ノード「更新履歴」に対応して抽出された領域

“1999.12.16　／　第1.1版”

を検索対象とする（Ｓ１０４）。そして、この記述パターン情報には繰り返し及び順序性が共に無しとして指定されているので、パターン検索部５３は、親ノー

ド「更新履歴」に対応する領域の先頭から、検索を開始する（Ｓ１０８，Ｓ１０９）。この検索において、この記述パターン情報にはパターン指定方法が開始及び終了と指定され、開始パターンが正規表現による“《領域の先頭》”、終了パターンが文字列そのものによる“／”と指定されているので、

“1999.12.16”

の部分が、記述パターン情報に合致した領域として検出される。従って、この領域が、ノード「更新日付」に対応した領域として抽出され、出力結果ツリー５６に追加される（Ｓ１１１）。

【００８２】

このノード「更新日付」には子ノードがなく（Ｓ１１２）、その記述パターン情報には繰り返しが無しと指定されているので（Ｓ１１５）、全体コントロール部５２は、その次弟ノード「版数」を、新たな処理対象ノードとする（Ｓ１１６，Ｓ１１７）。そして、全体コントロール部５２は、ＤＴＤ＋パターンツリー５５から、このノード「版数」についての記述パターン情報を参照して（Ｓ１０３）、親ノード「更新情報」に対応して抽出された領域

“1999.12.16　／　第1.1版”

を検索対象とする（Ｓ１０４）。そして、この記述パターン情報には繰り返しが無しと指定されているが順序性が有りと指定されているので、パターン検索部５３は、兄ノード「更新日付」の後から、検索を開始する（Ｓ１０６，Ｓ１０９）。この検索において、この記述パターン情報にはパターン指定方法が開始及び終了と指定され、開始パターンが文字列そのものによる“／”、終了パターンが正規表現による“《行末》”、と指定されているので、

“第1.1版”

の部分が、記述パターン情報に合致した領域として検出される。従って、この領域が、ノード「版数」に対応した領域として抽出され、出力結果ツリー５６に追加される（Ｓ１１１）。

【００８３】

このノード「版数」には子ノードがなく（Ｓ１１２）、その記述パターン情報には繰り返しが無しと指定されており（Ｓ１１５）、しかも弟ノードがないので

（Ｓ１１６）、全体コントロール部５２は、親ノード「更新履歴」を仮の処理対象ノードとする（Ｓ１１８）。当該仮の処理対象ノード「更新履歴」の記述パターン情報には繰り返し有りと指定されいるので（Ｓ１１５）、全体コントロール部５２は、このノード「更新履歴」による領域抽出を繰り返す。この場合、処理が２回目となるので、全体コントロール部５２は、親ノード「開発履歴」に対応する領域（処理対象文書Ｔの全体）における前回の処理で抽出された領域

“1999.12.16　／　第1.1版”

の後から、検索を開始する（Ｓ１０７，Ｓ１０９）。この検索においては、

“2000.02.14　／　第1.2版”

の部分が、記述パターン情報に合致した領域として最初に検出される。また、その後で行われるノード「更新日付」，ノード「版数」についての検索では、

“2000.02.14”

“第1.2版”

の部分が、夫々検出される。

【００８４】

その後においては、全体コントロール部５２は、再度、ノード「更新履歴」についての検索を試みるが、もはや記述パターンに合致する領域は検出されず（Ｓ１１０）、このノード「更新履歴」には弟ノードもないので、全体コントロール部５２は、一旦、親ノード「開発履歴」を仮の処理対象ノードとし（Ｓ１１８）、当該処理対象ノード「開発履歴」がルートノードであるが故に（Ｓ１１９）、検索及び出力結果ツリー５５の作成を終了する。この時点におけるＤＴＤ＋パターンツリー５５は、図２２に示す通りになる。

【００８５】

全体コントロール部５２は、このＤＴＤ＋パターンツリー５５に基づいて、その各ノードに付された文字列に対して、タグを付加することによって、図２３に示すような構造化文書を作成して出力する（Ｓ００８）。

（構造文書化システムの詳細構成及び処理内容）

次に、上述したＤＴＤ＋パターン作成支援システム７による処理内容を、詳細に説明する。図２７は、このＤＴＤ＋パターン作成支援システム７の処理内容（

ＤＴＤ＋パターン作成支援システム７を構築するプログラムによるＣＰＵ１の処理内容）を示すフローチャートである。

【００８６】

このＤＴＤ＋パターン作成支援システム７は、入力装置８を介してオペレータによって入力される開始コマンドによって起動し、図２４に示すような選択画面をディスプレイ４上に表示するとともに、この選択画面における各典型パターン選択ボタン７４１に、対応する典型パターン定義情報Ｓを関連付ける。そして、入力装置８を介してオペレータによる入力によって処理対象文書Ｔのサンプルが選択されると、ＤＴＤ＋パターン作成支援システム７は、この処理対象文書Ｔのサンプルをハードディスク２からＲＡＭ３上に読み込み、選択画面におけるサンプルプレビューボックス７２にそのテキスト内容を表示する。そして、オペレータが、このサンプルプレビューボックス７２内に表示されているテキストにおける何れかの行がドラッグすることによって選択した後で、その行のパターンに最も似ている典型パターンを探し出して、その典型パターンに対応した典型パターン選択ボタン７４１をクリックすると、ＤＴＤ＋パターン作成支援システム７は、図２７の処理を開始する。

【００８７】

図２７の処理では、ＤＴＤ＋パターン作成支援システム７は、スタート後最初のＳ２０１において、オペレータが選択した行を、ＲＡＭ３上の作業領域に読み込む。

【００８８】

次のＳ２０２では、ＤＴＤ＋パターン作成支援システム７は、オペレータがクリックした典型パターン選択ボタン７４１に関連付けられた典型パターン定義情報Ｓを、ＲＡＭ３上の作業領域に読み込む。そして、読み込んだ典型パターン定義情報Ｓの構造特定部Ｓ１に記載されている典型パターンの概略構造を、分解する。具体的には、典型パターンの概略構造における各エレメント（隅付括弧で囲まれた部分を、他の部分から抽出する。

【００８９】

次のＳ２０３では、ＤＴＤ＋パターン作成支援システム７は、Ｓ２０２にて分

解された各エレメント（隅付括弧で囲まれた部分）を先頭から一つづつ検索対象として特定し、各検索対象エレメントに関して文字種特定部Ｓ２にて特定されている正規表現パターンに合致した部分を、Ｓ２０１にてＲＡＭ３上の作業領域に読み込まれたテキストから探し出す。このとき、ＤＴＤ＋パターン作成支援システム７は、最初のエレメントを検索対象とする場合には、Ｓ２０１にてＲＡＭ３上の作業領域に読み込まれたテキストの先頭から検索を行い、それ以降のエレメントを検索対象とする場合には、直前のエレメントに関して探し出された部分の後から検索を行う。

【００９０】

次のＳ２０４では、ＤＴＤ＋パターン作成支援システム７は、図２８に示すようなダイアログ７００を、選択画面に重ねて表示する。このダイアログ７００は、典型パターン定義情報Ｓ毎に作成される。図２８に示す例は、図２５に示される典型パターン定義情報Ｓに関連付けて作成されたものであるので、要素名テキストボックス７０１，タイトル文字列テキストボックス７０２，タイトル文字列リストボックス７０３，区切り文字列テキストボックス７０４，区切り文字列リストボックス７０５，追加ボタン７０６が、含まれている。ＤＴＤ＋パターン作成支援システム７は、Ｓ２０３にて各エレメントに関して探し出された部分を、対応するテキストボックス７０２，７０４に表示する。

【００９１】

図２８は、サンプルプレビューボックス７２内に表示されたテキスト中の

“会社名：富士通株式会社”

の行が選択された後に、図２５に示す典型パターン情報Ｓに関連付けられた典型パターン選択ボタン７４１がクリックされた場合を示しているので、エレメント［タイトルパターン］に関して探し出された部分“会社名”がタイトル文字列テキストボックス７０２に表示されているとともに、エレメント［区切りパターン］に関して探し出された部分“：”が区切り文字列テキストボックス７０４に表示されている。

【００９２】

なお、タイトル文字列リストボックス７０３には、オペレータが、タイトル文

字列テキストボックス７０２内に表示されている文字列と置換可能な文字列を入力し得る。同様に、区切り文字列リストボックス７０５には、オペレータが、区切り文字列テキストボックス７０４内に表示されている文字列と置換可能な文字列を入力し得る。また、要素名テキストボックス７０１には、オペレータが、作成しようとする記述パターンが指定される要素の要素名を入力し得る。そして、追加ボタン７０６をオペレータがクリックすると、ＤＴＤ＋パターン作成支援システム７は、処理をＳ２０５に進める。

【００９３】

Ｓ２０５では、ＤＴＤ＋パターン作成支援システム７は、典型パターン情報Ｓ中の雛形部Ｓ３における各記述パターン情報の雛形の［］内に、ダイアログ７００の各欄に表示された文字列を、雛形部Ｓ３内で指定された表現（文字種特定部Ｓ２で特定された表現よりは具体的な表現）に変換してから代入する。図２５及び図２６の例では、タイトル文字列テキストボックス７０２内に表示されている文字列が正規表現に変換されて［タイトル文字列１］に代入され、タイトル文字列リストボックス７０３内に表示されている文字列が正規表現に変換されて［タイトル文字列２］に代入され、区切り文字列テキストボックス７０４内に表示されている文字列が正規表現に変換されて［区切り文字列１］に代入され、区切り文字列リストボックス７０５内に表示されている文字列が［区切り文字列２］に代入される。これにより、雛形部Ｓ３内の雛形が、要素名テキストボックス７０１内に表示されている要素名を有する要素に対して指定される記述パターン情報となって、ＤＴＤ＋パターン情報Ｒに追加される。図２９は、図２８に示す状態で追加ボタン７０６がクリックされた場合に作成される記述パターン情報を示す。なお、上述したように、この時点において、ツリーリストボックス７３内には、図３０に示すように、ルートノード「設計書」の子ノードとして、要素名「会社名」が表示される。

【００９４】

以後、オペレータが、サンプルプレビューボックス７３内に表示されているテキスト中の任意の行を選択して、何れかの典型パターン選択ボタン７４１をクリックする毎に、新たな子ノード（若しくは、子ノード及び孫ノード）についての

記述パターン情報が作成されて、ＤＴＤ＋パターン情報Ｒに追加される。

【００９５】

例えば、ルートノード「設計書」の子ノードとして要素「会社情報」が追加された状態において、サンプルプレビューボックス７２内に表示されたテキスト中の

“ファイル名＜日本語名＞ファイル長

KOKYAKU-MASTER＜顧客マスタ＞200”

の行が選択された後に、図２６に示す典型パターン情報Ｓに関連付けられた典型パターン選択ボタン７４１がクリックされた場合のダイアログ７００’を、図３１に示す。このダイアログ７００’には、タイトル文字列テキストボックス７０２が５個含まれ、区切り文字列テキストボックス７０４が４個含まれている。また、追加ボタン７０６の代わりに、ＯＫボタン７０７が含まれている。

【００９６】

このＯＫボタン７０７がクリックされると、図２６に示す典型パターン情報Ｓ中の各雛形部Ｓ３中の［］に、ダイアログ７００’の各欄に表示された文字列に基づいて変換された文字列が、代入される。その結果として、ツリーリストボックス７３内には、図３２に示すように、ルートノード「設計書」の子ノート及び孫ノードとして、要素名「ファイル属性」等が表示される。

【００９７】

【発明の効果】

以上説明したように、文書構造の各要素の抽出条件の内容を任意に設定しておき、テキスト形式で記述された処理対象電子文書にこの抽出条件を当てはめることによって、文書構造の各要素に該当する領域を抽出することができるので、抽出した各領域にその要素に対応したタグを付加することよって、自動的に構造化文書を生成することができる。

【図面の簡単な説明】

【図１】　本発明の実施の形態である構造文書化システムの概念をＤＴＤ＋パターン編集システム及びＤＴＤ＋パターン作成支援システムの概念とともに示す概念図

【図２】　構造文書化システム等が実現されるコンピュータの構成を示すブロック図

【図３】　構造文書化システムの詳細なモジュール構成を示すプログラム構成部

【図４】　構造文書化システムによる処理手順を示すフローチャート

【図５】　図４のＳ００７にて実行される出力結果ツリー作成サブルーチンを示すフローチャート

【図６】　図４のＳ００７にて実行される出力結果ツリー作成サブルーチンを示すフローチャート

【図７】　ＤＴＤ＋パターンツリーの構成例を示す図

【図８】　処理対象文書のテキスト例を示す図

【図９】　出力結果ツリーの構成例を示す図

【図１０】　構造化文書例を示す図

【図１１】　正規化表現のルールを示す表

【図１２】　ＤＴＤ＋パターンツリーの構成例を示す図

【図１３】　処理対象文書のテキスト例を示す図

【図１４】　ＢＮＦの定義の一部を示す表

【図１５】　構文要素の範囲を示す図

【図１６】　構文木・コメント木の構成例を示す図

【図１７】　出力結果ツリーの構成例を示す図

【図１８】　ＤＴＤ＋パターン編集システムによる編集画面例を示す図

【図１９】　ＤＴＤ＋パターン情報のテキスト例を示す図

【図２０】　ＤＴＤ＋パターンツリーの構成例を示す図

【図２１】　処理対象文書のテキスト例を示す図

【図２２】　出力結果ツリーの構成例を示す図

【図２３】　構造化文書例を示す図

【図２４】　ＤＴＤ＋パターン作成支援システムによる選択画面例を示す図

【図２５】　典型パターン定義情報のテキスト例を示す図

【図２６】　典型パターン定義情報のテキスト例を示す図

【図２７】　ＤＴＤ＋パターン作成支援システムによる処理手順を示すフローチャート

【図２８】　ＤＴＤ＋パターン作成支援システムによる選択画面例を示す図

【図２９】　ＤＴＤ＋パターン作成支援システムによって作成される記述パターン例を示す図

【図３０】　ＤＴＤ＋パターン作成支援システムによる選択画面例を示す図

【図３１】　ＤＴＤ＋パターン作成支援システムによる選択画面例を示す図

【図３２】　ＤＴＤ＋パターン作成支援システムによる選択画面例を示す図

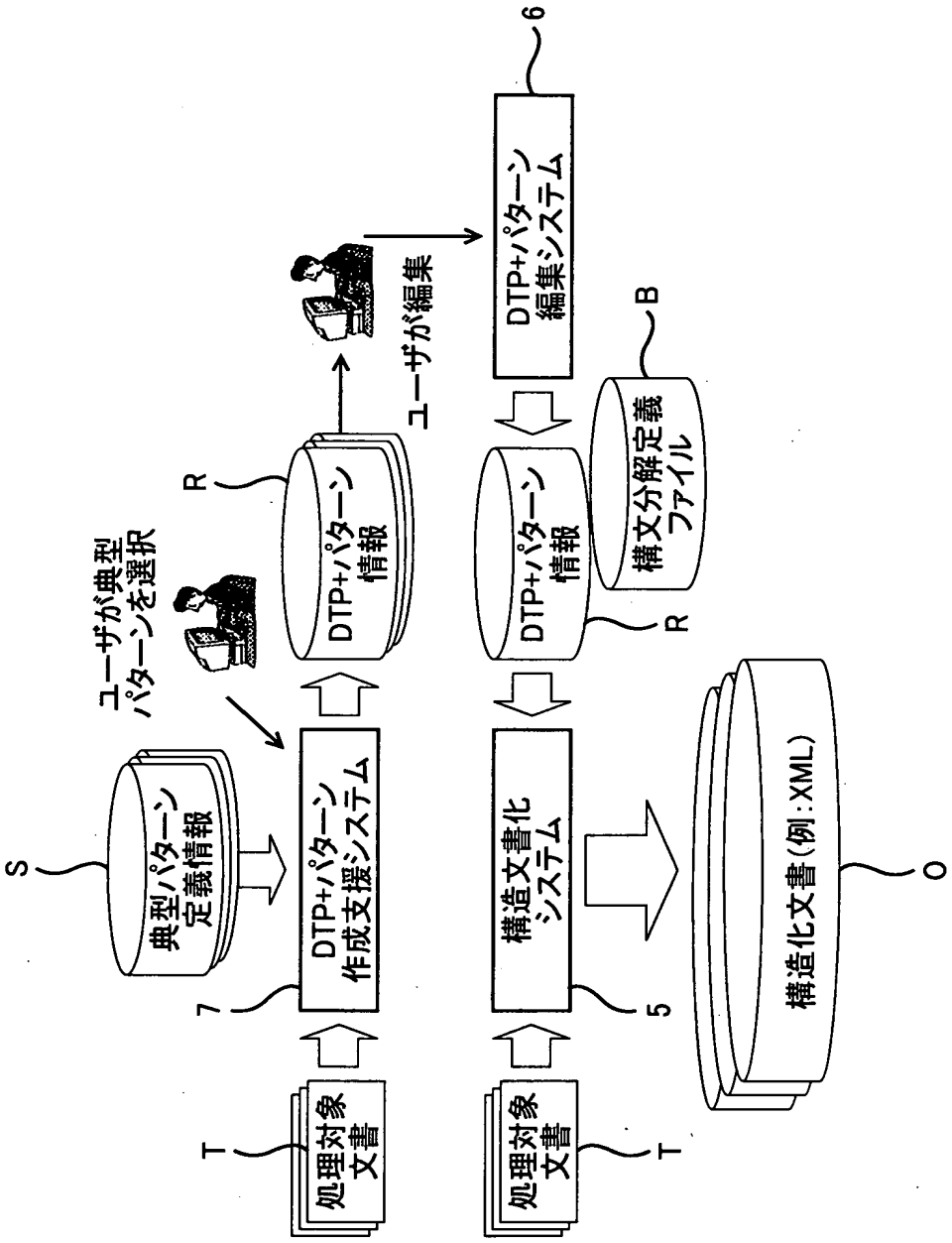
【図３３】　従来の構造化文書の説明図

【符号の説明】

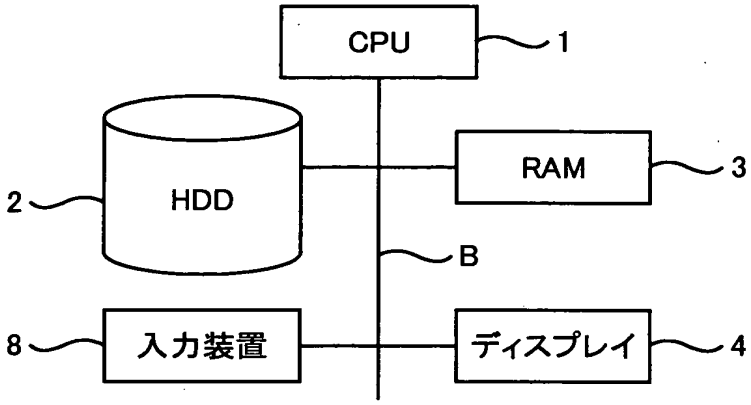
１	ＣＰＵ
２	ハードディスク
３	ＲＡＭ
５	構造文書化システム
８	入力装置
５２	全体コントロール部
５３	検索部
５４	構文木分解部
Ｒ	ＤＴＤ＋パターン情報
Ｔ	処理対象文書
Ｏ	構造化文書

【書類名】　　図面

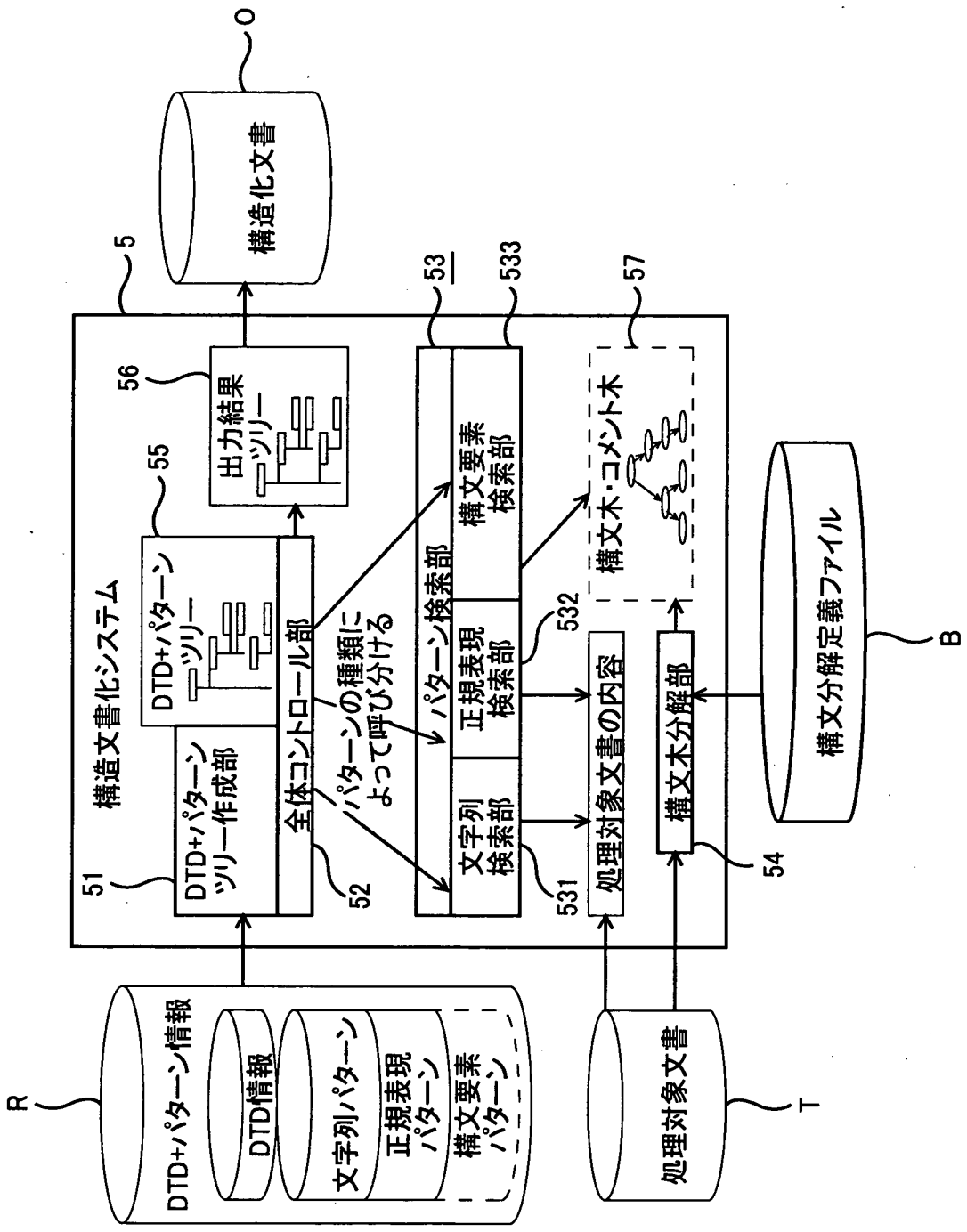
【図１】


S
典型パターン
定義情報
ユーザが典型
パターンを選択
7
DTP+パターン
作成支援システム
R
DTP+パターン
情報
ユーザが編集
6
DTP+パターン
編集システム
B
構文分解定義
ファイル
DTP+パターン
情報
R
T
処理対象
文書
構造文書化
システム
5
構造化文書(例:XML)
O
T
処理対象
文書

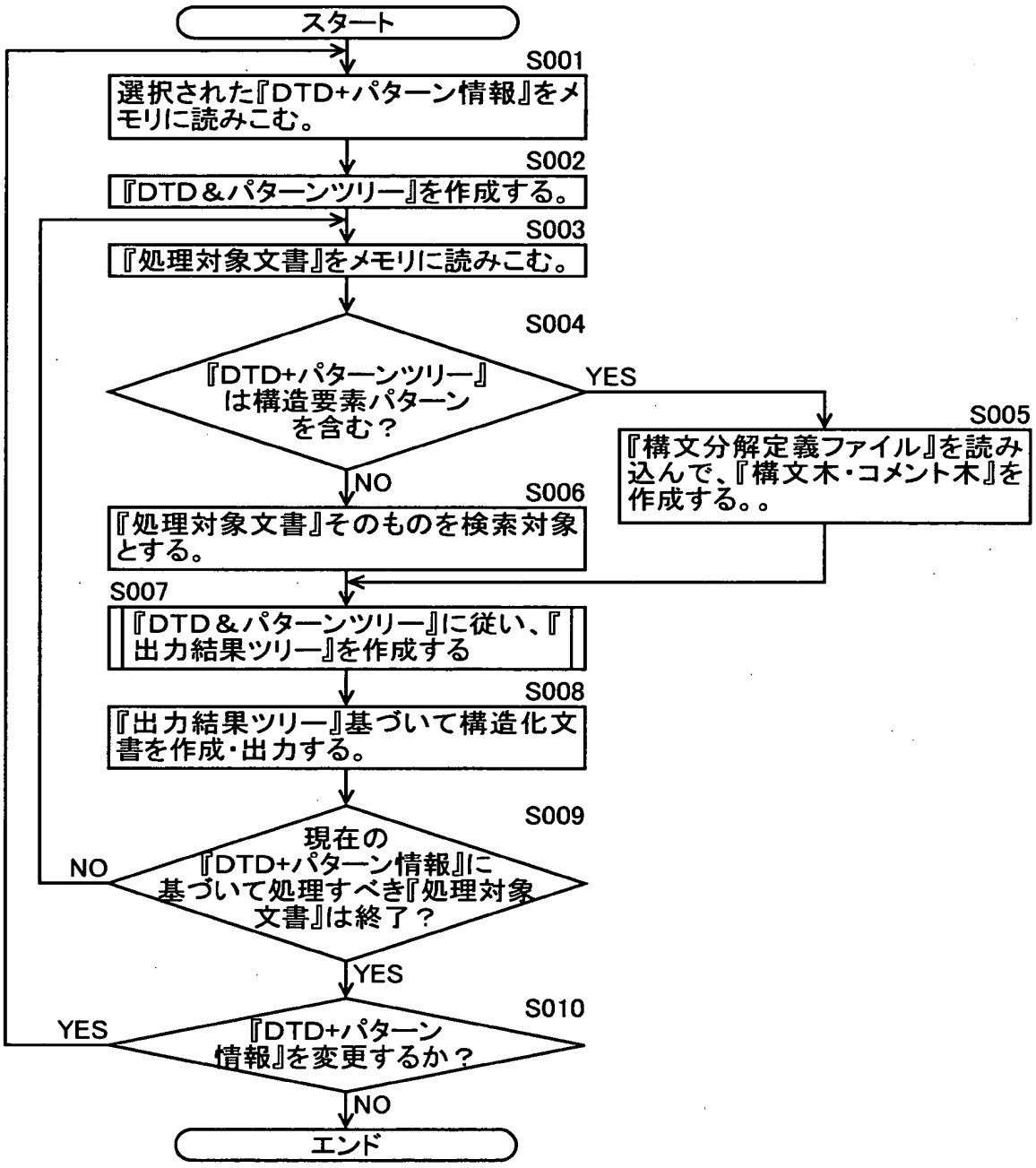
【図２】



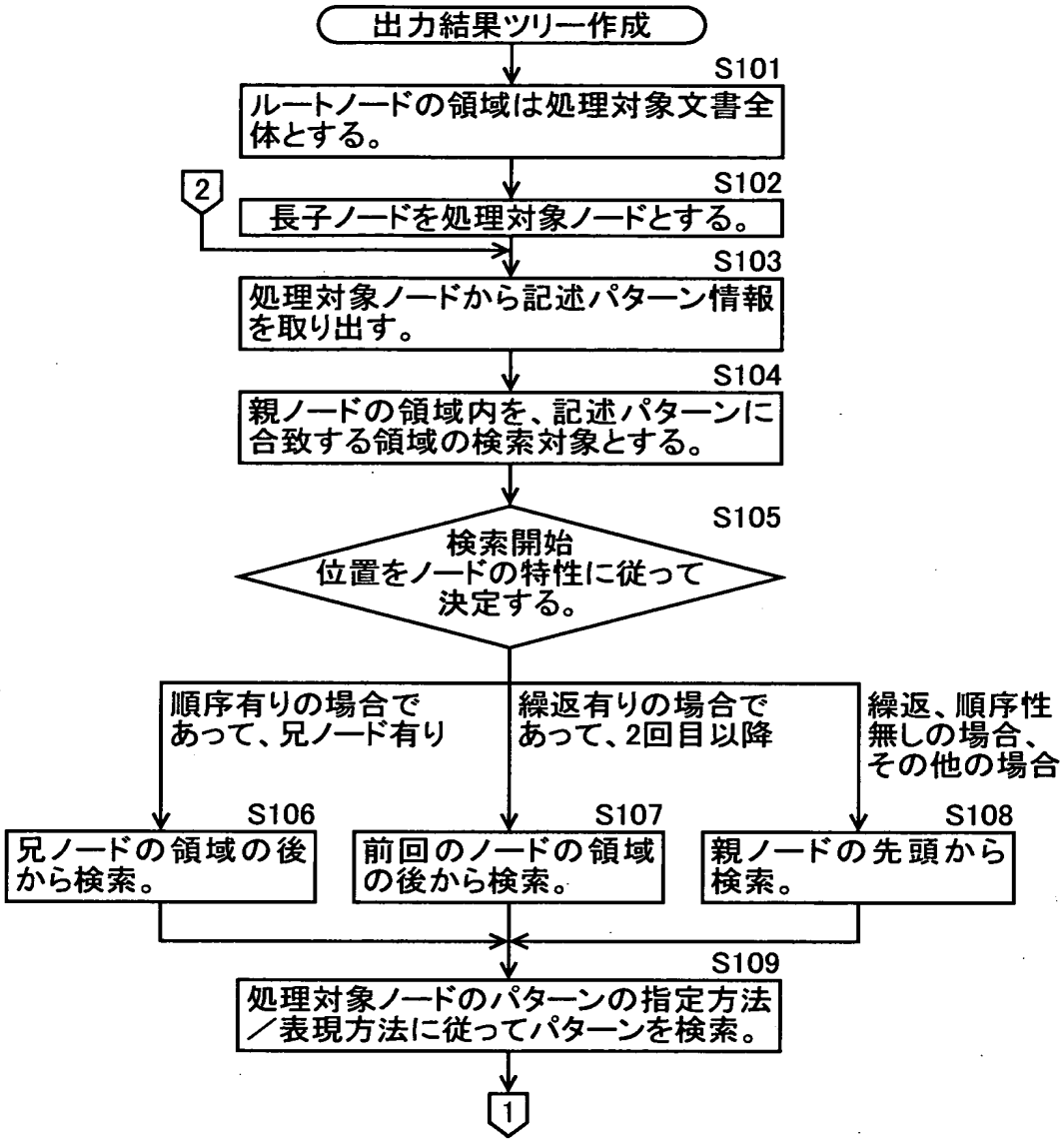
【図３】


構造文書化システム
5
DTD+パターン情報
DTD情報
文字列パターン
正規表現パターン
構文要素パターン
R
DTD+パターンツリー作成部
51
全体コントロール部
52
DTD+パターンツリー
55
出力結果ツリー
56
構造化文書
0
パターンの種類によって呼び分ける
パターン検索部
53
文字列検索部
531
正規表現検索部
532
構文要素検索部
533
処理対象文書の内容
構文木・コメント木
57
構文木分解部
54
構文分解定義ファイル
B
処理対象文書
T

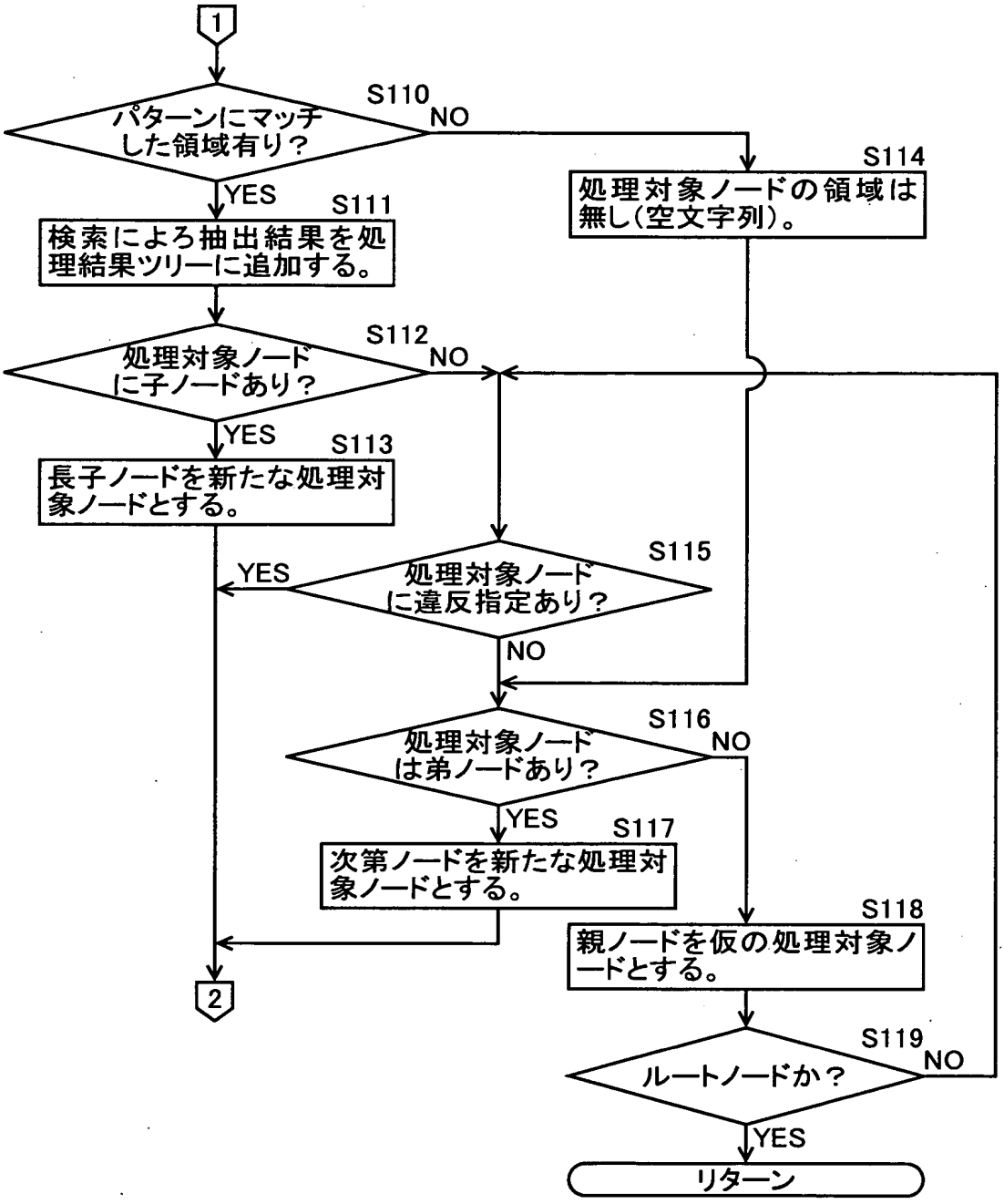
【図４】


スタート
S001
選択された『DTD+パターン情報』をメモリに読みこむ。
S002
『DTD＆パターンツリー』を作成する。
S003
『処理対象文書』をメモリに読みこむ。
S004
『DTD+パターンツリー』は構造要素パターンを含む？
YES
S005
『構文分解定義ファイル』を読み込んで、『構文木・コメント木』を作成する。。
NO
S006
『処理対象文書』そのものを検索対象とする。
S007
『DTD＆パターンツリー』に従い、『出力結果ツリー』を作成する
S008
『出力結果ツリー』基づいて構造化文書を作成・出力する。
S009
現在の『DTD+パターン情報』に基づいて処理すべき『処理対象文書』は終了？
NO
YES
S010
『DTD+パターン情報』を変更するか？
YES
NO
エンド

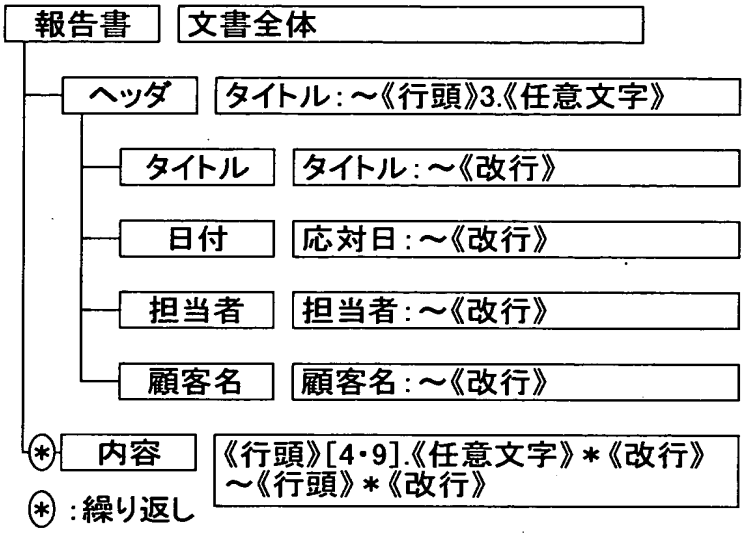
【図５】


出力結果ツリー作成
S101
ルートノードの領域は処理対象文書全体とする。
2
S102
長子ノードを処理対象ノードとする。
S103
処理対象ノードから記述パターン情報を取り出す。
S104
親ノードの領域内を、記述パターンに合致する領域の検索対象とする。
S105
検索開始位置をノードの特性に従って決定する。
順序有りの場合であって、兄ノード有り
繰返有りの場合であって、2回目以降
繰返、順序性無しの場合、その他の場合
S106
兄ノードの領域の後から検索。
S107
前回のノードの領域の後から検索。
S108
親ノードの先頭から検索。
S109
処理対象ノードのパターンの指定方法／表現方法に従ってパターンを検索。
1

【図６】


1
S110
パターンにマッチした領域有り？
NO
YES
S111
検索によろ抽出結果を処理結果ツリーに追加する。
S112
処理対象ノードに子ノードあり？
NO
YES
S113
長子ノードを新たな処理対象ノードとする。
S114
処理対象ノードの領域は無し(空文字列)。
S115
処理対象ノードに違反指定あり？
YES
NO
S116
処理対象ノードは弟ノードあり？
NO
YES
S117
次第ノードを新たな処理対象ノードとする。
S118
親ノードを仮の処理対象ノードとする。
S119
ルートノードか？
NO
YES
リターン
2

【図７】



【図８】

○○部長殿

タイトル：商談対応報告書
担当者：藤川　泰之

1.顧客名　山田証券

2.応対日：1997.02.17

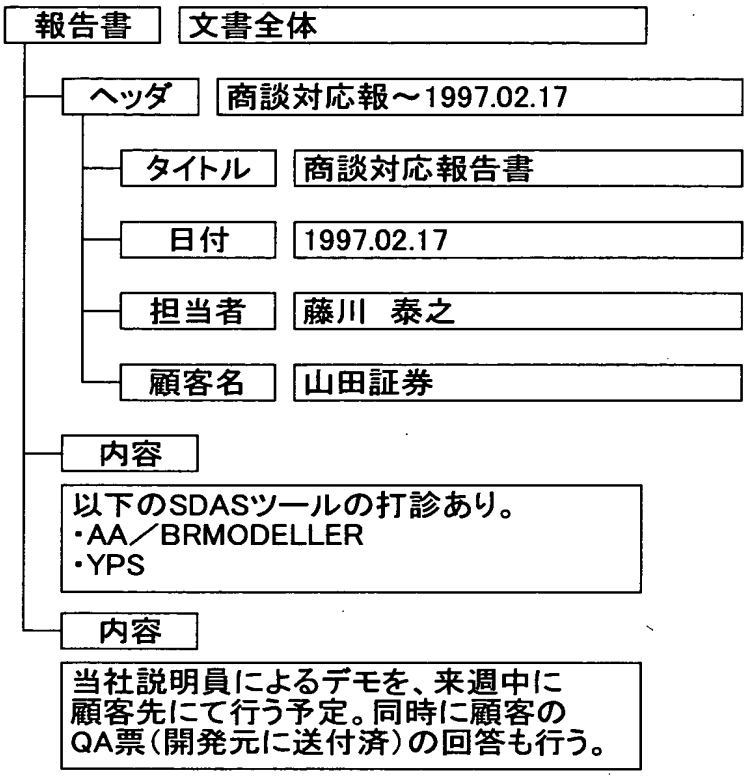
3.概要
　上記顧客より当社ツール使用の打診。

4.概要
以下のSDASツールの打診あり。
・AA／BRMODELLER
・YPS

5.今後の予定
当社説明員によるデモを、来週中に
顧客先にて行う予定。同時に顧客の
QA票（開発元に送付済）の回答も行う。

以上

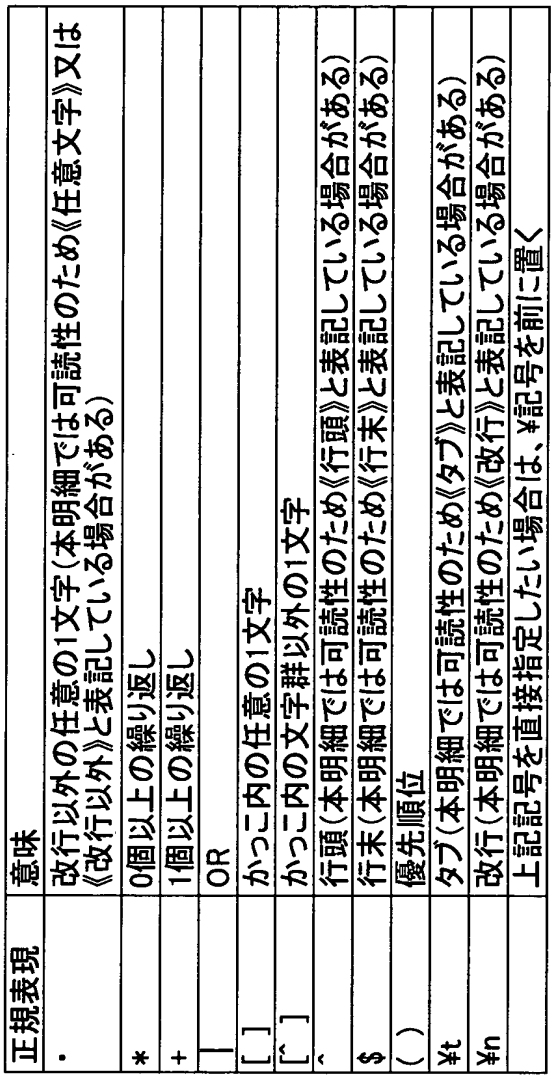
【図９】



【図１０】

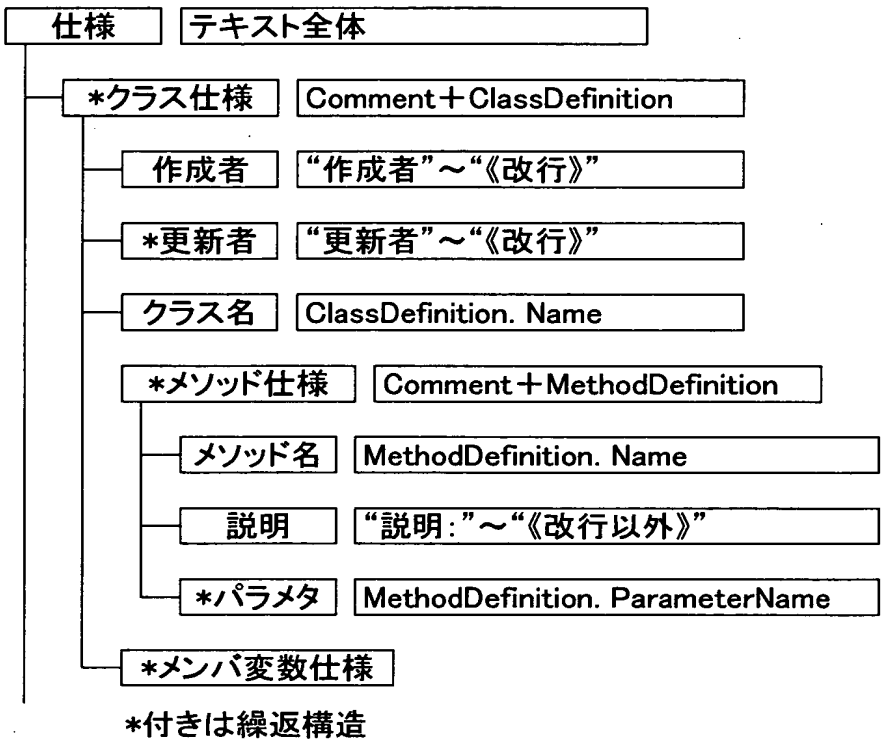
```
<報告書>
<ヘッダ>
<タイトル>商談対応報告書</タイトル>
<日付>1997.02.17</日付>
<担当者>藤川　泰之</担当者>
<顧客名>山田証券</顧客名>
</ヘッダ>
<内容>
以下のSDASツールの打診あり。
・AA／BRMODELLER
・YPS
</内容>
<内容>
当社説明員によるデモを、来週中に
顧客先にて行う予定。同時に顧客の
QA票（開発元に送付済）の回答も行う。
</内容>
<報告書>
```

【図１１】



正規表現	意味
.	改行以外の任意の1文字（本明細では可読性のため《任意文字》又は《改行以外》と表記している場合がある）
*	0個以上の繰り返し
+	1個以上の繰り返し
|	OR
[ ]	かっこ内の任意の1文字
[^]	かっこ内の文字群以外の1文字
^	行頭（本明細では可読性のため《行頭》と表記している場合がある）
$	行末（本明細では可読性のため《行末》と表記している場合がある）
( )	優先順位
¥t	タブ（本明細では可読性のため《タブ》と表記している場合がある）
¥n	改行（本明細では可読性のため《改行》と表記している場合がある）
	上記記号を直接指定したい場合は、¥記号を前に置く

【図１２】



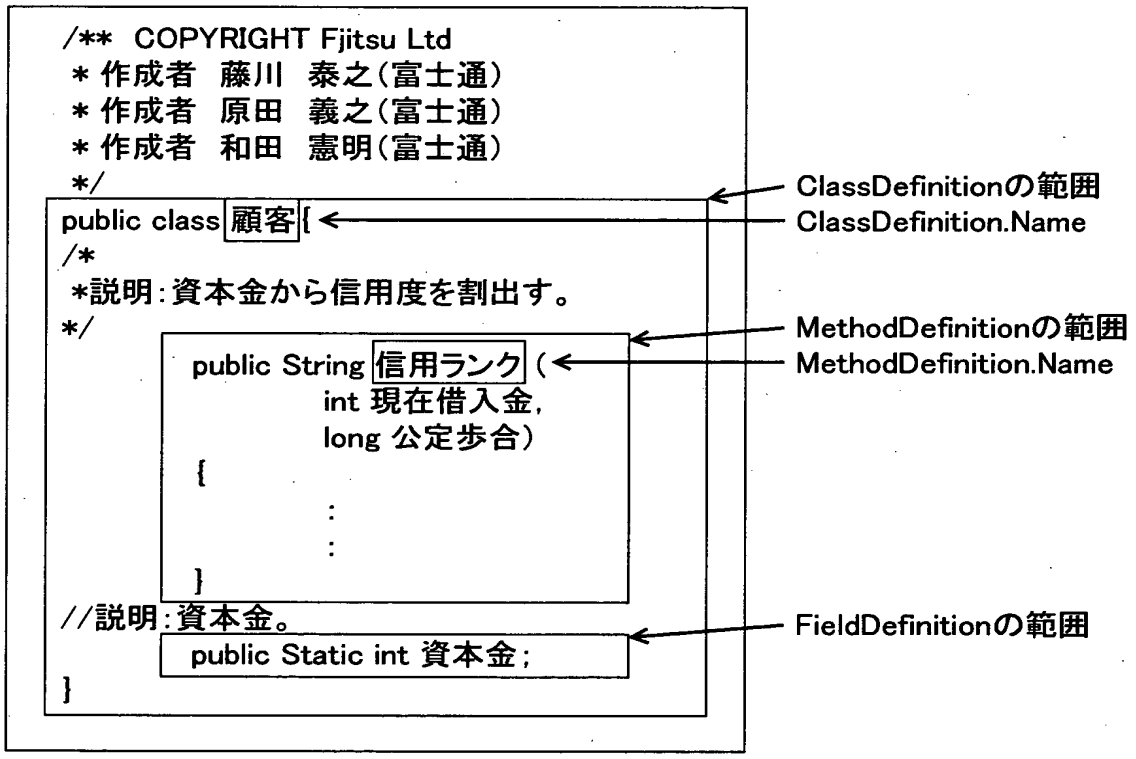
【図１３】

```
/**  COPYRIGHT Fjitsu Ltd
 * 作成者  藤川  泰之(富士通)
 * 更新者  原田  義之(富士通)
 * 更新者  和田  憲明(富士通)
 */
public class 顧客{
/*
 *説明:資本金から信用度を割出す。
*/
        public String 信用ランク(
                int 現在借入金,
                long 公定歩合)
        {
                :
                :
        }
//説明:資本金。
        public Static int 資本金;
}
```

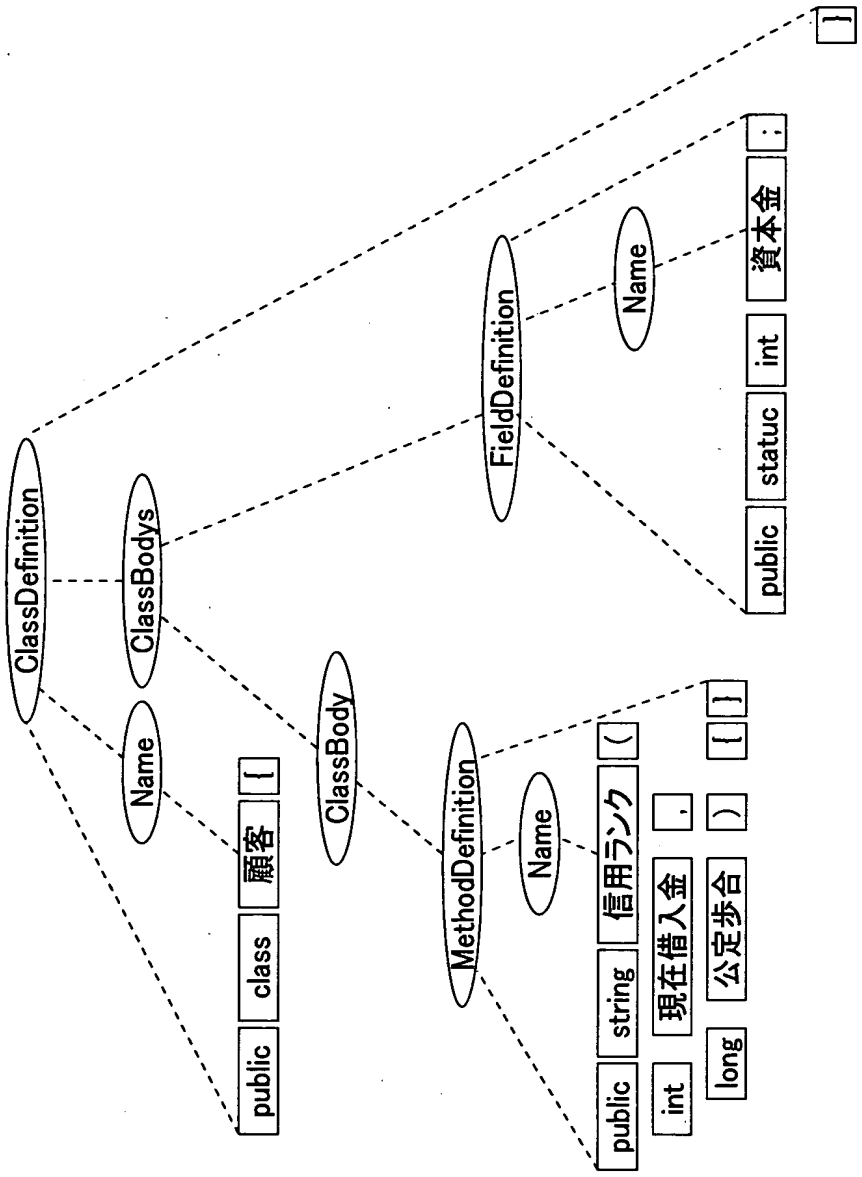
【図１４】

```
ClassDefinition:"public""class"Name"{"ClassBodys"}";
ClassBodys:ClassBody";"
        | ClassBodys ClassBodys";";
ClassBody:FieldDefinition
        | MethodDefinition;
MethodDefinition:"public"Type Name '(' FormalParamenterList '(' '{' Block '}';
FormalParameterList:FormalParameter
        | FormalParameterList ',' FormalParameter;
FormalParameter:TypeParameterName;
                                    :
                                    :
```

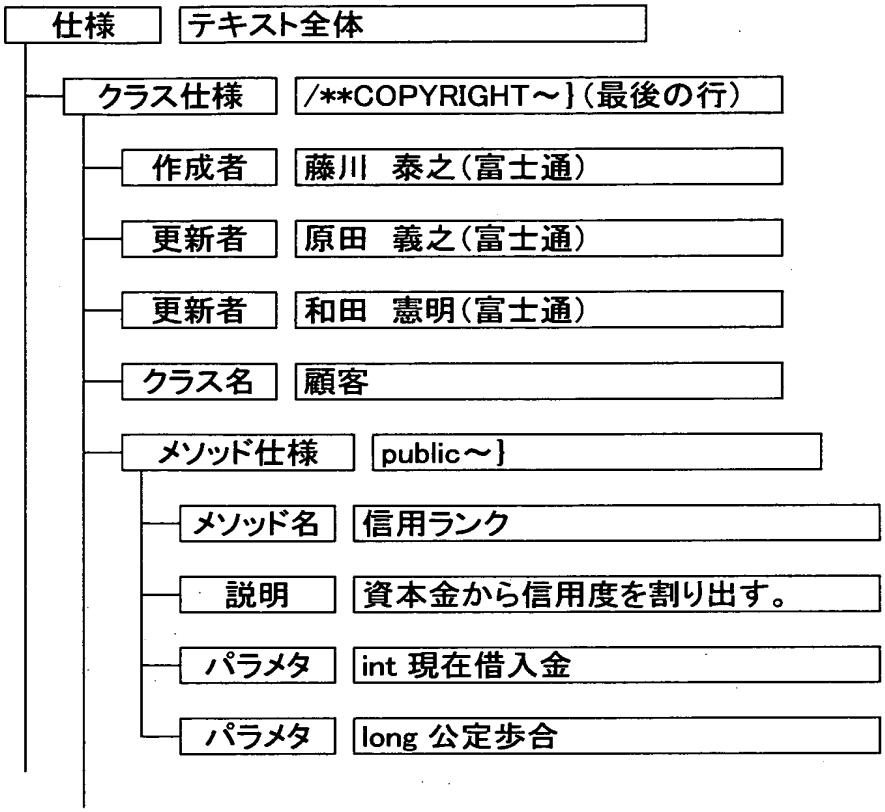
【図１５】



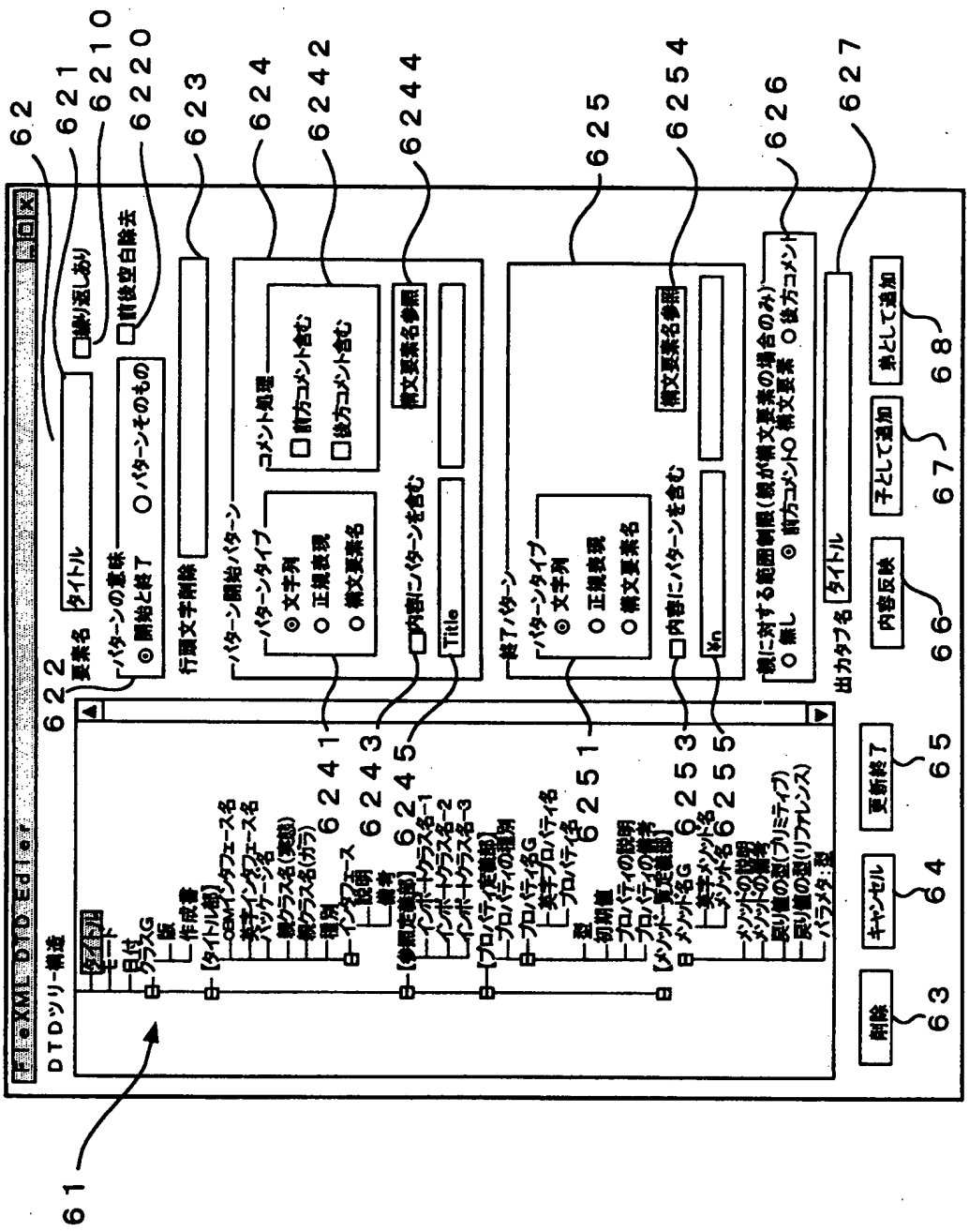
【図１６】


ClassDefinition
Name
public
class
顧客
{
ClassBodys
ClassBody
FieldDefinition
public
statuc
int
資本金
;
Name
MethodDefinition
public
string
信用ランク
(
int
現在借入金
,
long
公定歩合
)
{
}
Name
}

【図１７】


仕様
テキスト全体
クラス仕様
/**COPYRIGHT～}（最後の行）
作成者
藤川　泰之（富士通）
更新者
原田　義之（富士通）
更新者
和田　憲明（富士通）
クラス名
顧客
メソッド仕様
public～}
メソッド名
信用ランク
説明
資本金から信用度を割り出す。
パラメタ
int 現在借入金
パラメタ
long 公定歩合

【図１８】


XML DTD Editor
DTDツリー構造
繰り返しあり
前後空白除去
タイトル
パターンの意味
◎ 開始と終了
○ パターンそのもの
行頭文字削除
パターン開始パターン
パターンタイプ
◎ 文字列
○ 正規表現
○ 構文要素名
コメント処理
□ 前方コメント含む
□ 後方コメント含む
構文要素名参照
□ 内容にパターンを含む
Title
終了パターン
パターンタイプ
◎ 文字列
○ 正規表現
○ 構文要素名
構文要素名参照
□ 内容にパターンを含む
\n
親に対する範囲制限（親が構文要素の場合のみ）
○ 無し ◎ 前方コメント○ 構文要素 ○ 後方コメント
出力タブ名
タイトル
内容反映
子として追加
弟として追加
更新終了
キャンセル
削除
62
621
6210
6220
623
624
6242
6244
625
6254
626
627
622
6241
6243
6245
6251
6253
6255
66
67
68
65
64
63
61

【図１９】

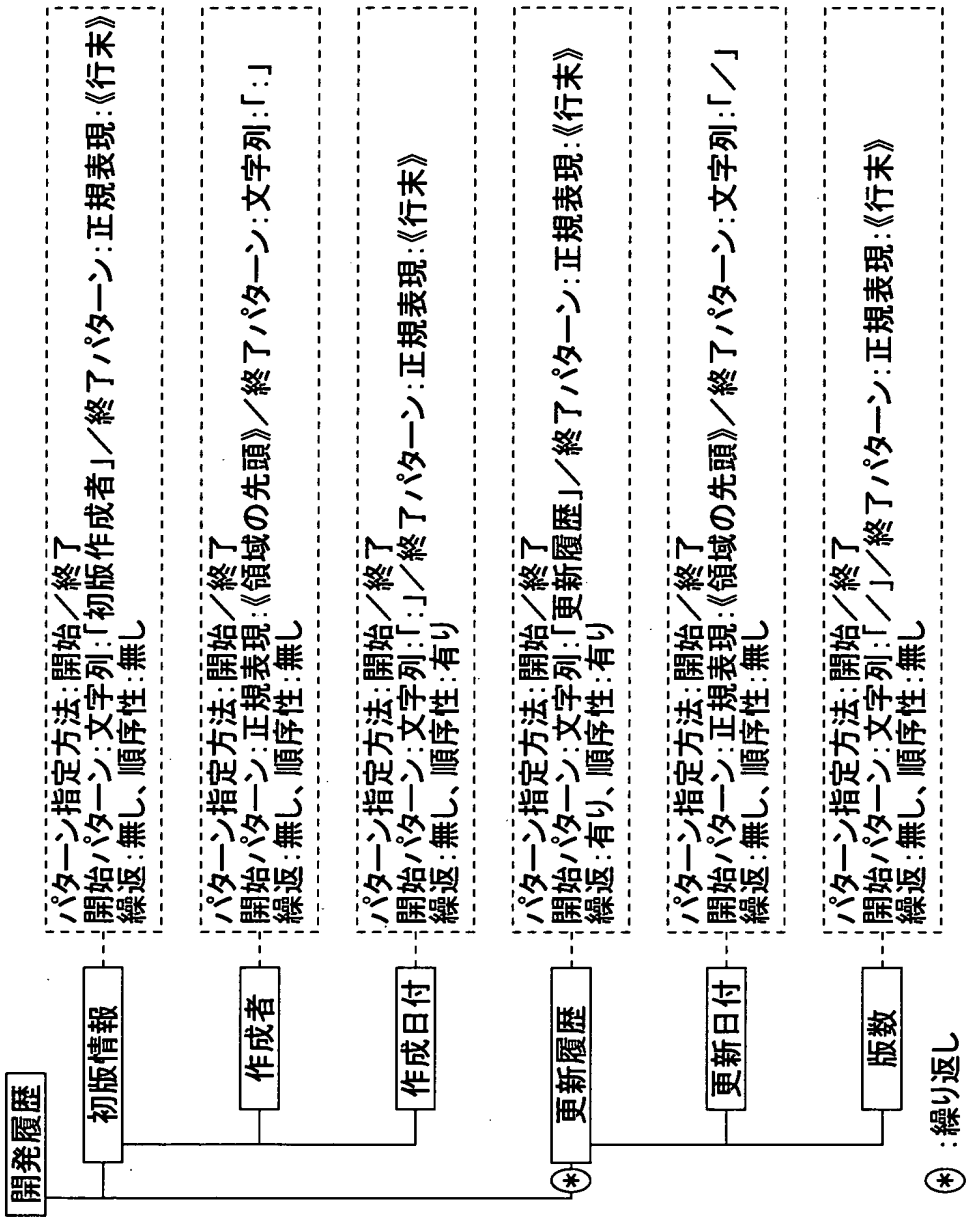
[開発履歴]
ルート:YES
子ノード＝初版情報、更新履歴

[初版情報]
子ノード＝作成者、作成日付
パターン指定方法:開始／終了
開始パターン:文字列:「初版作成者」
終了パターン:正規表現:《行末》
繰返:無し
順序性:無し

[作成者]
パターン指定方法:開始／終了
開始パターン:正規表現:《領域の先頭》
終了パターン:文字列:「:」
繰返:無し
順序性:無し
[作成日付]
パターン指定方法:開始／終了
開始パターン:文字列:「:」
終了パターン:正規表現:《行末》
繰返:無し
順序性:有り

[更新履歴]
子ノード＝更新日付、版数
パターン指定方法:開始／終了
開始パターン:文字列:「更新履歴」
終了パターン:正規表現:《行末》
繰返:有り
順序性:有り
[更新日付]
パターン指定方法:開始／終了
開始パターン:正規表現:《領域の先頭》
終了パターン:文字列:「／」
繰返:無し
順序性:無し
[版数]
パターン指定方法:開始／終了
開始パターン:文字列:「／」
終了パターン:正規表現:《行末》
繰返:無し
順序性:無し

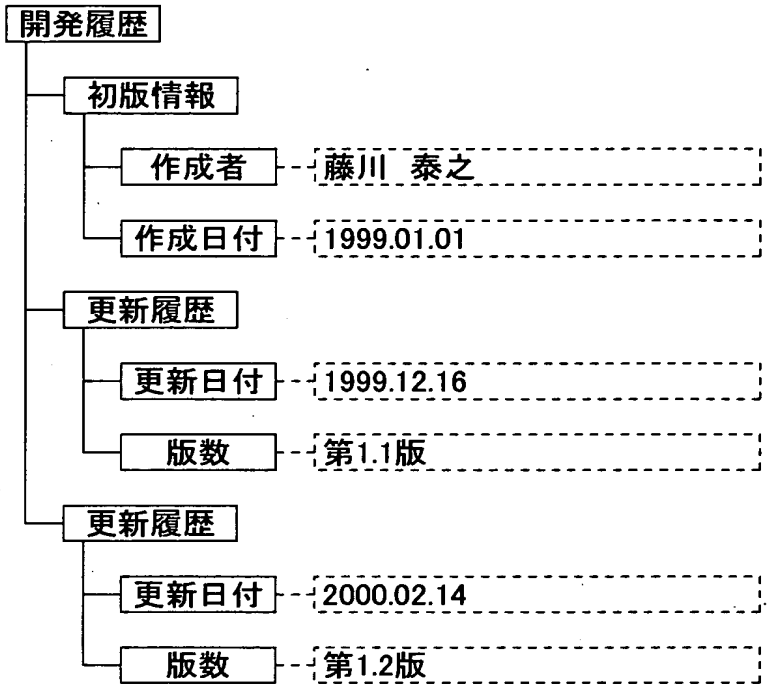
【図２０】



【図２１】

```
＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊
初版作成者　藤川 泰之 ： 1999.01.01
＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊
更新履歴　1999.12.16 ／ 第1.1版
更新履歴　2000.02.14 ／ 第1.2版
＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊
```

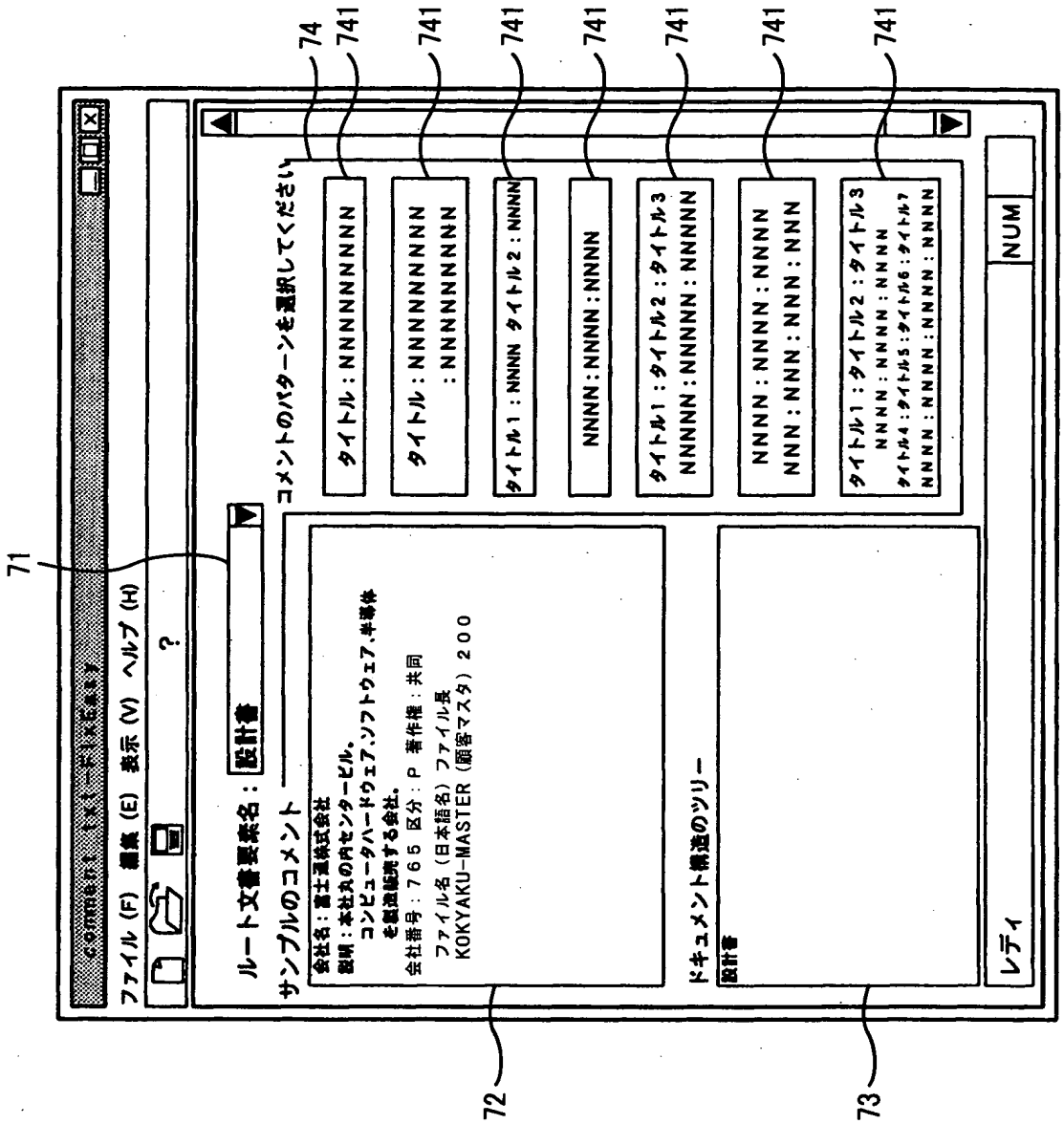
【図２２】



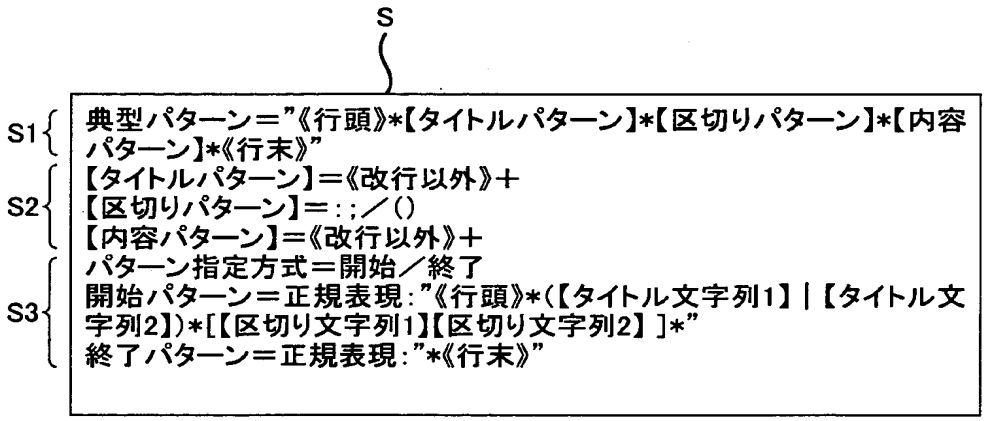
【図２３】

```
<開発履歴>
<初版情報>
<作成者>藤川　泰之</作成者>
<作成日付>1999.01.01</作成日付>
<更新履歴>
<更新日付>1999.12.16</更新日付>
<版数>第1.1版</版数>
</更新履歴>
<更新履歴>
<更新日付>2000.02.14</更新日付>
<版数>第1.2版</版数>
</更新履歴>
</開発履歴>
```

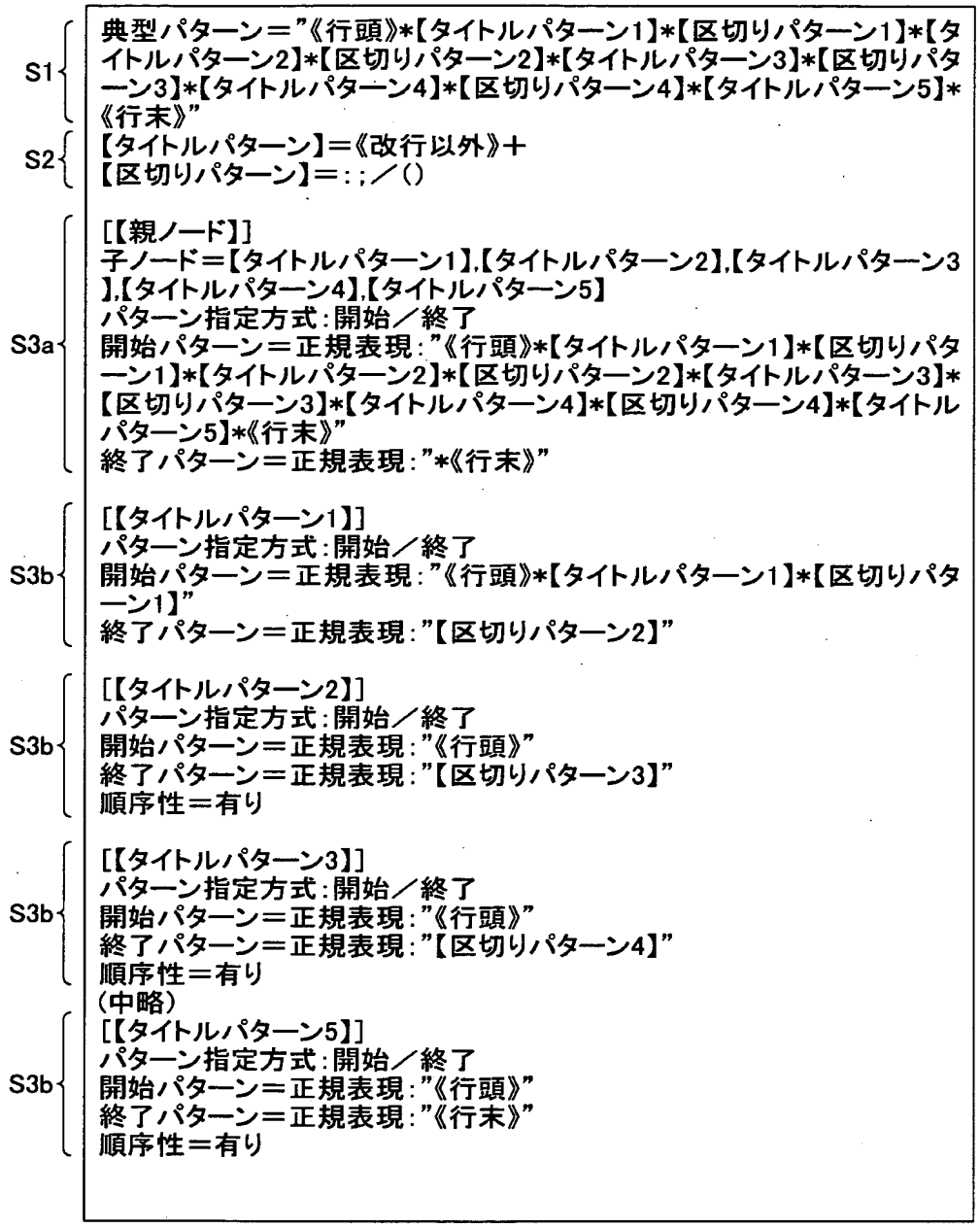
【図２４】


71
72
73
74
741
ファイル (F) 編集 (E) 表示 (V) ヘルプ (H)
ルート文書要素名：
設計書
サンプルのコメント
会社名：富士通株式会社
説明：本社丸の内センタービル、
コンピュータハードウェア、ソフトウェア、半導体
を製造販売する会社。
会社番号：７６５ 区分：Ｐ 著作権：共同
ファイル名（日本語名）ファイル長
KOKYAKU-MASTER（顧客マスタ）２００
コメントのパターンを選択してください
タイトル：NNNNNNNNN
タイトル：NNNNNNNNN
：NNNNNNNNN
タイトル１：NNNN タイトル２：NNNN
NNNN：NNNN：NNNN
タイトル１：タイトル２：タイトル３
NNNNN：NNNNN：NNNNN
NNNN：NNNN：NNNN
NNN：NNN：NNN：NNN
タイトル１：タイトル２：タイトル３
NNNN：NNNN：NNNN
タイトル４：タイトル５：タイトル６：タイトル７
NNNN：NNNN：NNNN：NNNN
ドキュメント構造のツリー
設計書
NUM
レディ

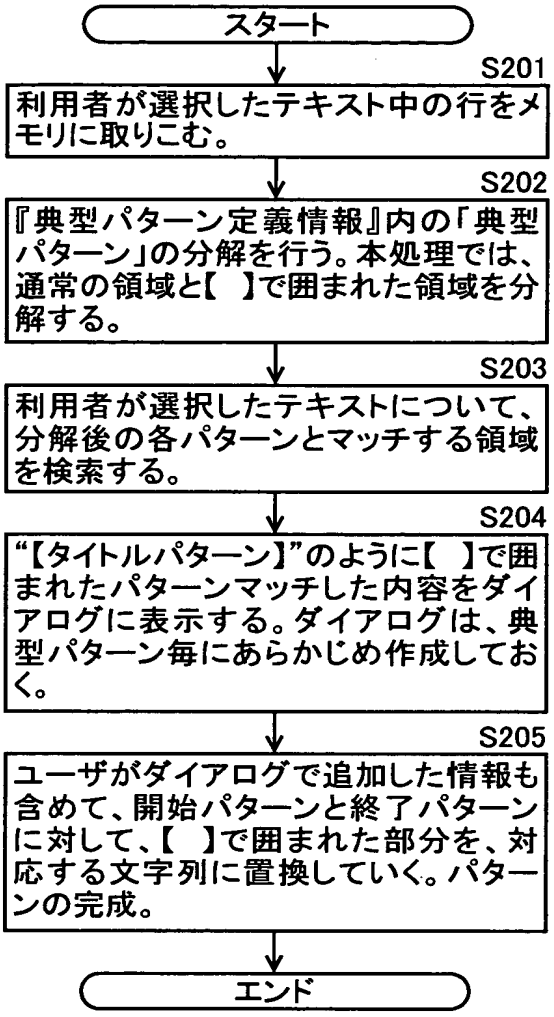
【図２５】


S
S1
典型パターン＝"《行頭》*【タイトルパターン】*【区切りパターン】*【内容パターン】*《行末》"
S2
【タイトルパターン】＝《改行以外》＋
【区切りパターン】＝：；／()
【内容パターン】＝《改行以外》＋
S3
パターン指定方式＝開始／終了
開始パターン＝正規表現："《行頭》*(【タイトル文字列1】|【タイトル文字列2】)*[【区切り文字列1】【区切り文字列2】]*"
終了パターン＝正規表現："*《行末》"

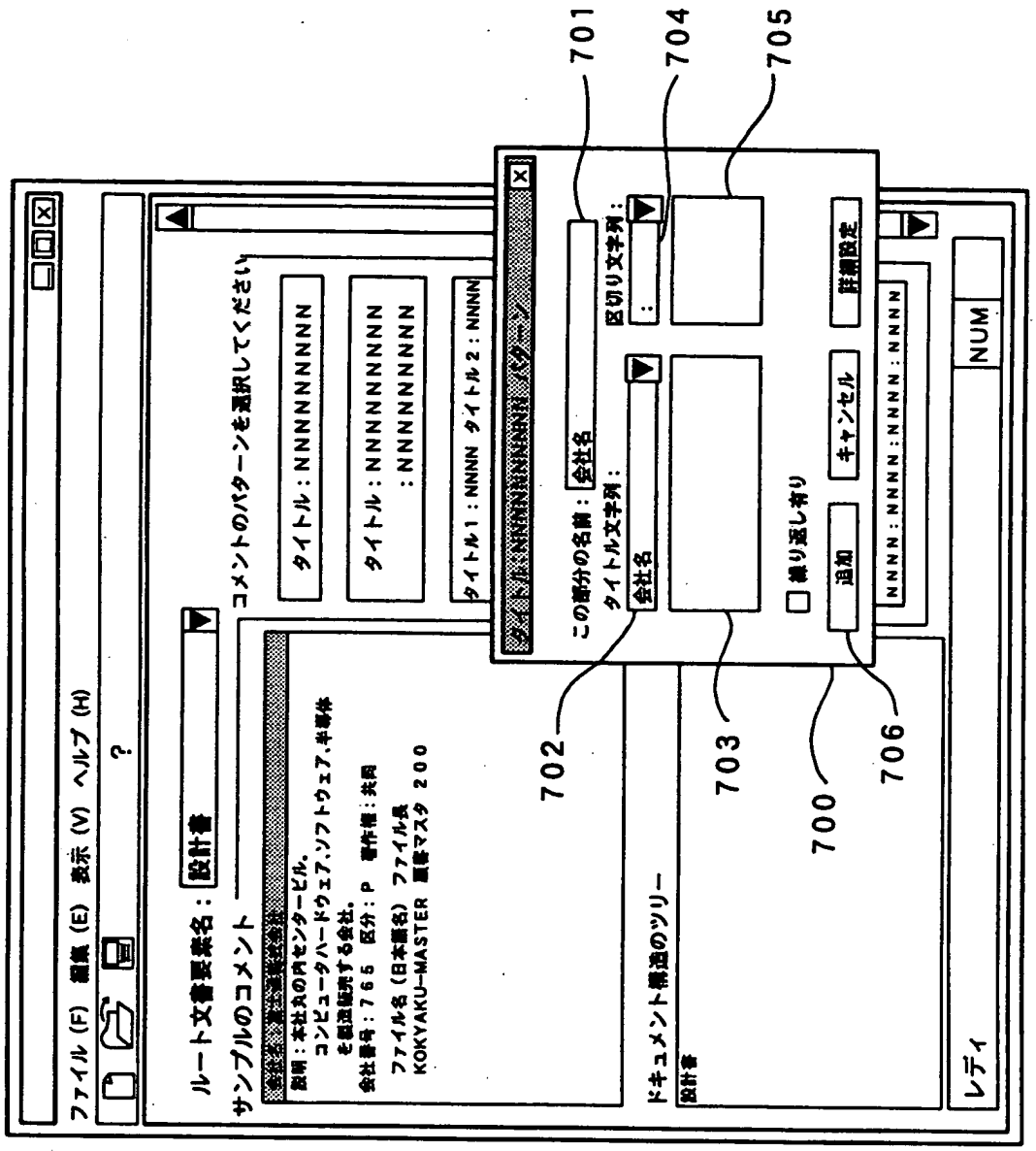
【図２６】


S1
典型パターン＝"《行頭》*【タイトルパターン1】*【区切りパターン1】*【タイトルパターン2】*【区切りパターン2】*【タイトルパターン3】*【区切りパターン3】*【タイトルパターン4】*【区切りパターン4】*【タイトルパターン5】*《行末》"
S2
【タイトルパターン】＝《改行以外》＋
【区切りパターン】＝：；／()
S3a
[【親ノード】]
子ノード＝【タイトルパターン1】,【タイトルパターン2】,【タイトルパターン3】,【タイトルパターン4】,【タイトルパターン5】
パターン指定方式：開始／終了
開始パターン＝正規表現："《行頭》*【タイトルパターン1】*【区切りパターン1】*【タイトルパターン2】*【区切りパターン2】*【タイトルパターン3】*【区切りパターン3】*【タイトルパターン4】*【区切りパターン4】*【タイトルパターン5】*《行末》"
終了パターン＝正規表現："*《行末》"
S3b
[【タイトルパターン1】]
パターン指定方式：開始／終了
開始パターン＝正規表現："《行頭》*【タイトルパターン1】*【区切りパターン1】"
終了パターン＝正規表現："【区切りパターン2】"
S3b
[【タイトルパターン2】]
パターン指定方式：開始／終了
開始パターン＝正規表現："《行頭》"
終了パターン＝正規表現："【区切りパターン3】"
順序性＝有り
S3b
[【タイトルパターン3】]
パターン指定方式：開始／終了
開始パターン＝正規表現："《行頭》"
終了パターン＝正規表現："【区切りパターン4】"
順序性＝有り
(中略)
S3b
[【タイトルパターン5】]
パターン指定方式：開始／終了
開始パターン＝正規表現："《行頭》"
終了パターン＝正規表現："《行末》"
順序性＝有り

【図２７】


スタート
S201
利用者が選択したテキスト中の行をメモリに取りこむ。
S202
『典型パターン定義情報』内の「典型パターン」の分解を行う。本処理では、通常の領域と【　】で囲まれた領域を分解する。
S203
利用者が選択したテキストについて、分解後の各パターンとマッチする領域を検索する。
S204
“【タイトルパターン】”のように【　】で囲まれたパターンマッチした内容をダイアログに表示する。ダイアログは、典型パターン毎にあらかじめ作成しておく。
S205
ユーザがダイアログで追加した情報も含めて、開始パターンと終了パターンに対して、【　】で囲まれた部分を、対応する文字列に置換していく。パターンの完成。
エンド

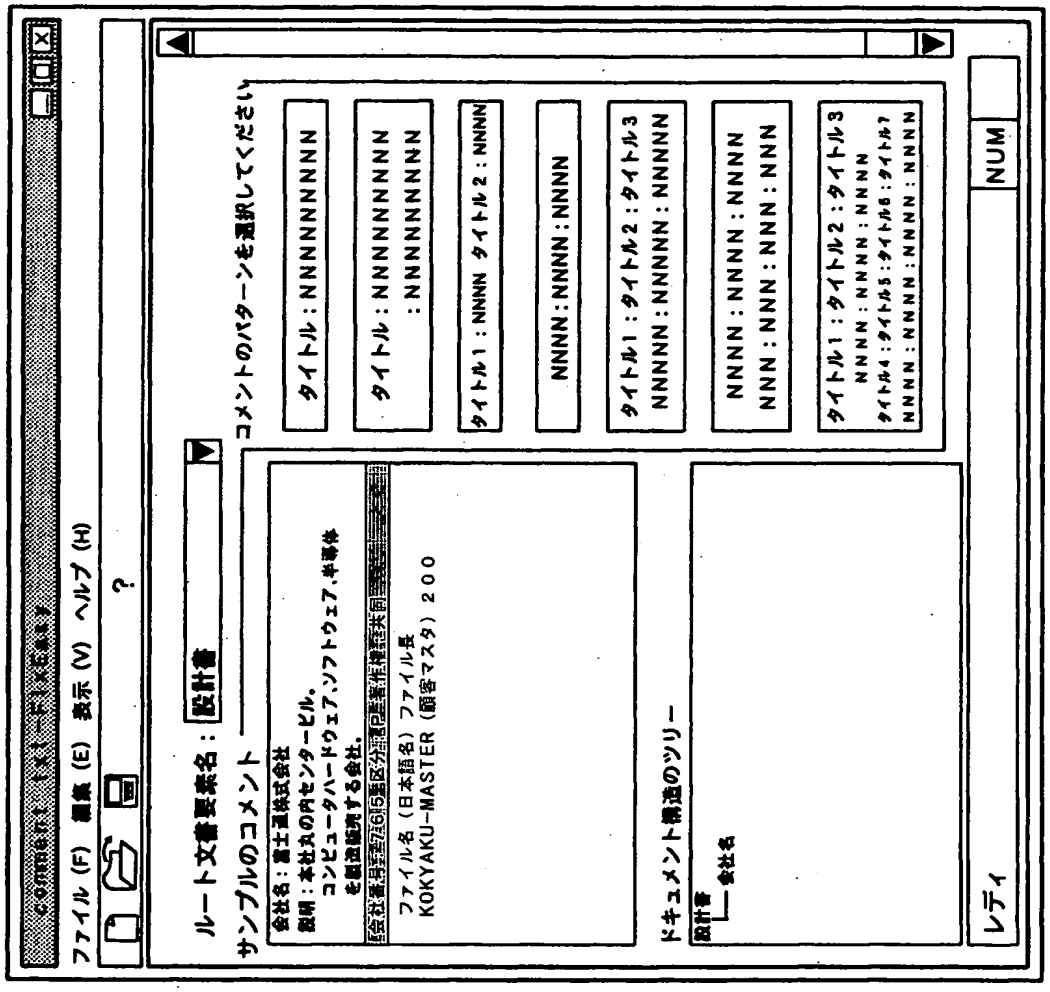
【図２８】


701
704
705
ファイル (F) 編集 (E) 表示 (V) ヘルプ (H)
ルート文書要素名：
設計書
コメントのパターンを選択してください
タイトル：NNNNNNNNN
タイトル：NNNNNNNNN
：NNNNNNNNN
タイトル1：NNNN タイトル2：NNNN
この部分の名前：
会社名
区切り文字列：
:
タイトル文字列：
会社名
繰り返し有り
追加
キャンセル
詳細設定
NNNN:NNNN:NNNN:NNNN
NUM
サンプルのコメント
説明：本社丸の内センタービル、
コンピュータハードウェア、ソフトウェア、半導体
を製造販売する会社。
会社番号：765 区分：P 著作権：共同
ファイル名（日本語名） ファイル属
KOKYAKU-MASTER 顧客マスタ 200
ドキュメント構造のツリー
設計書
レディ
702
703
700
706

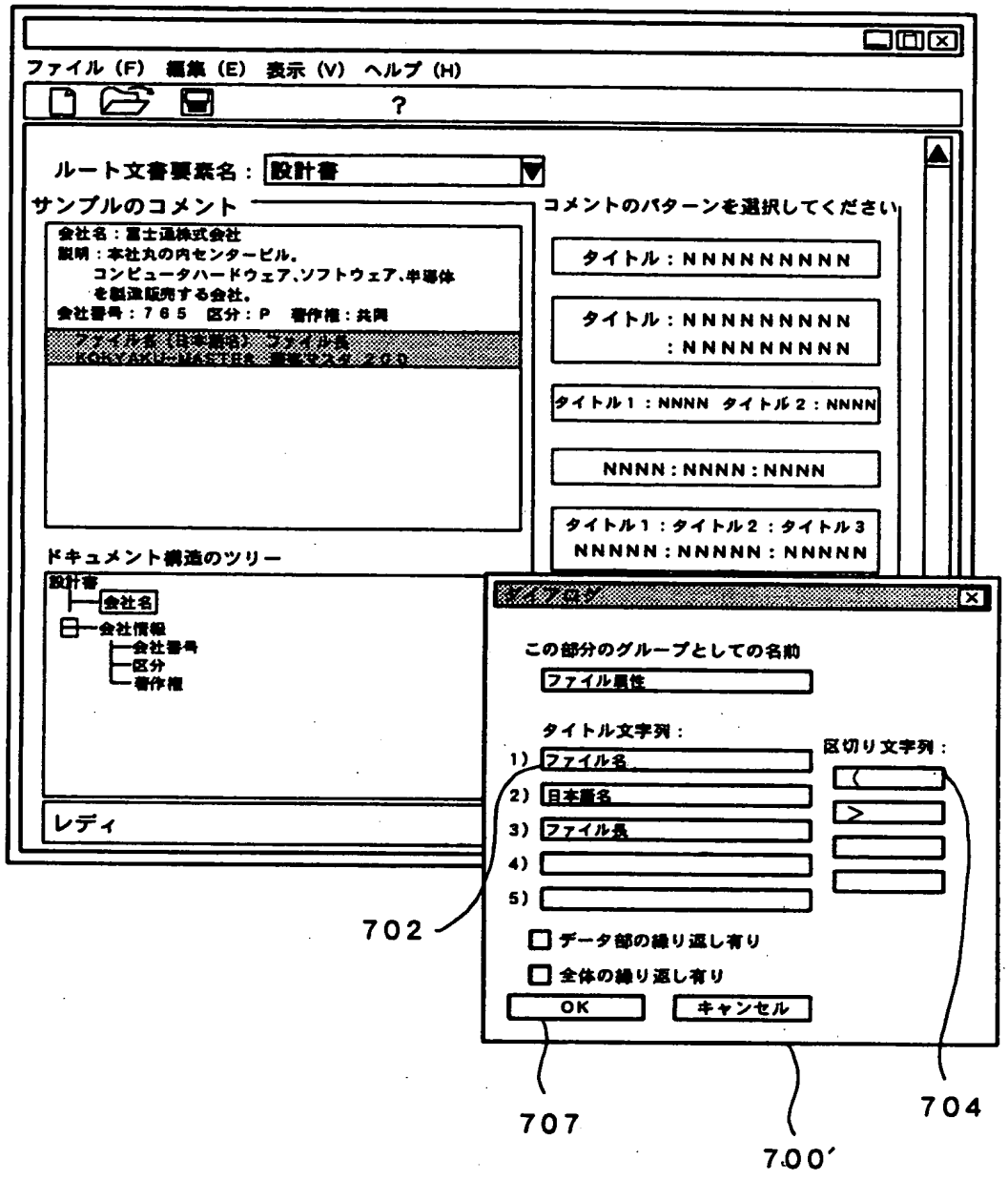
【図２９】

[会社名]
パターン指定方式:開始／終了
開始パターン＝正規表現:"《行頭》*(富士通)*[:]*"
終了パターン＝正規表現:"*《行末》"

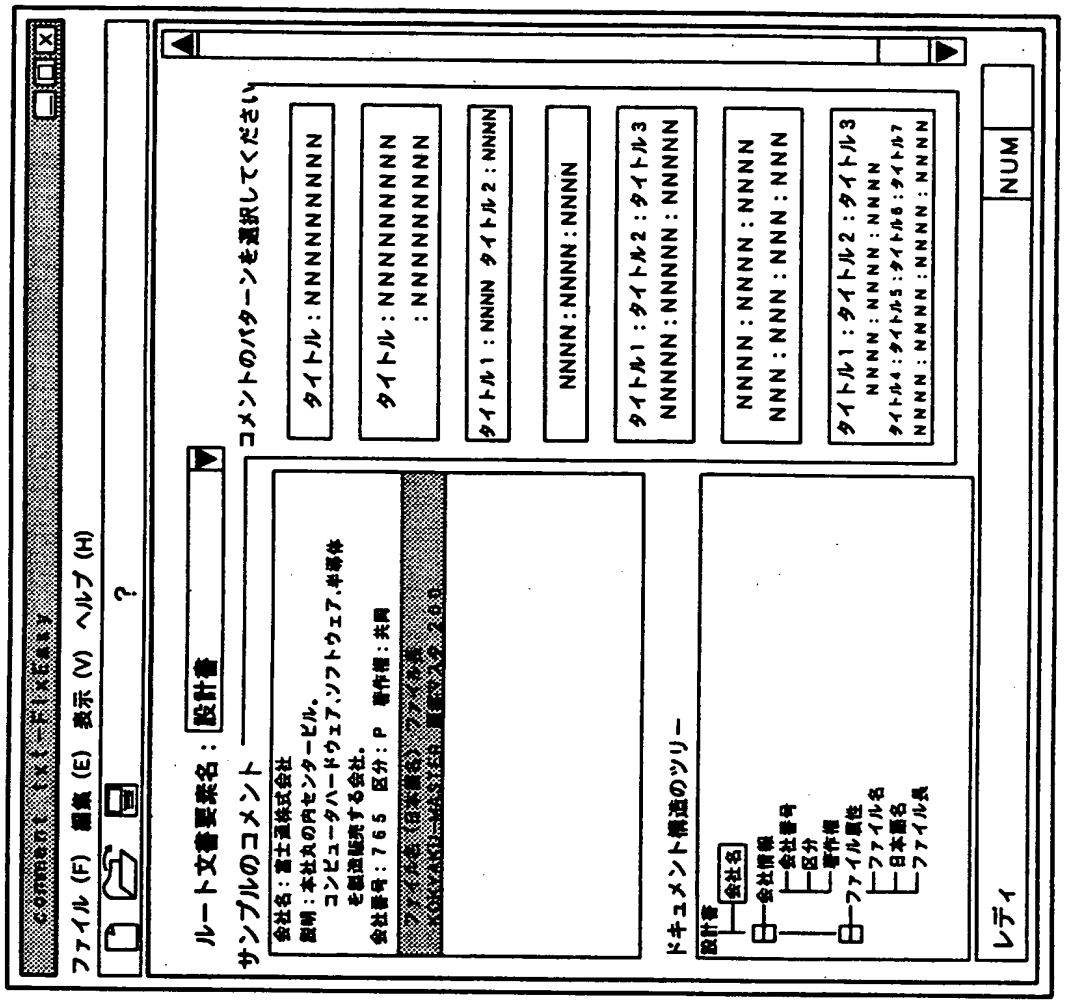
【図３０】



【図３１】


ファイル (F)　編集 (E)　表示 (V)　ヘルプ (H)
ルート文書要素名：設計書
サンプルのコメント
会社名：富士通株式会社
説明：本社丸の内センタービル。
コンピュータハードウェア、ソフトウェア、半導体
を製造販売する会社。
会社番号：７６５　区分：P　著作権：共同
コメントのパターンを選択してください
タイトル：NNNNNNNNN
タイトル：NNNNNNNNN
：NNNNNNNNN
タイトル１：NNNN　タイトル２：NNNN
NNNN：NNNN：NNNN
タイトル１：タイトル２：タイトル３
NNNNN：NNNNN：NNNNN
ドキュメント構造のツリー
設計書
会社名
会社情報
会社番号
区分
著作権
レディ
この部分のグループとしての名前
ファイル属性
タイトル文字列：
区切り文字列：
1) ファイル名
2) 日本語名
3) ファイル長
4)
5)
データ部の繰り返し有り
全体の繰り返し有り
OK
キャンセル
702
707
700′
704

【図３２】

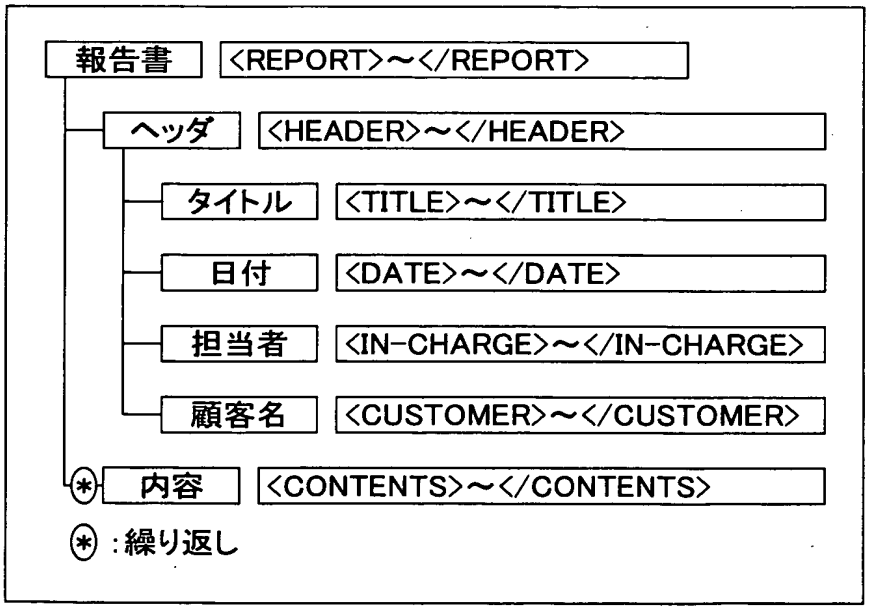

Comment Text Filter
ファイル (F) 編集 (E) 表示 (V) ヘルプ (H)
?
ルート文書要素名：
設計書
コメントのパターンを選択してください
サンプルのコメント
会社名：富士通株式会社
説明：本社大内センタービル、
コンピュータハードウェア、ソフトウェア、半導体
を製造販売する会社。
会社番号：７６５ 区分：Ｐ 著作権：共同
ファイル名（日本語名） ファイル長
ドキュメント構造のツリー
設計書
会社名
会社情報
会社番号
区分
著作権
ファイル属性
ファイル名
日本語名
ファイル長
タイトル：ＮＮＮＮＮＮＮＮＮ
タイトル：ＮＮＮＮＮＮＮＮＮ
：ＮＮＮＮＮＮＮＮＮ
タイトル１：ＮＮＮＮ タイトル２：ＮＮＮＮ
ＮＮＮＮ：ＮＮＮＮ：ＮＮＮＮ
タイトル１：タイトル２：タイトル３
ＮＮＮＮＮ：ＮＮＮＮＮ：ＮＮＮＮＮ
ＮＮＮＮ：ＮＮＮＮ：ＮＮＮＮ
ＮＮＮ：ＮＮＮ：ＮＮＮ：ＮＮＮ
タイトル１：タイトル２：タイトル３
ＮＮＮＮ：ＮＮＮＮ：ＮＮＮＮ
タイトル４：タイトル５：タイトル６：タイトル７
ＮＮＮＮ：ＮＮＮＮ：ＮＮＮＮ：ＮＮＮＮ
レディ
NUM

【図３３】

(a)
構造化テキスト(XML)の例

```
<REPORT>
<HEADER>
<TITLE>商談対応報告書</TITLE>
<DATE>1997.02.17</DATE>
<IN-CHARGE>藤川　泰之</IN-CHARGE>
<CUSTOMER>山田証券</CUSTOMER>
</HEADER>
<CONTENTS>
以下のSDASツールの打診あり。
・AA／BRMODELLER
・YPS
</CONTENTS>
<CONTENTS>
当社説明員によるデモを、来週中に
顧客先にて行う予定。同時に顧客の
QA票(開発元に送付済)の回答も行う。
</CONTENTS>
<REPORT>
```

(b)
文書構造定義(DTD)を図にしたもの



【書類名】　　要約書

【要約】

【課題】　テキストから自動的に構造化文書を生成することができる構造文書化システムを、提供する。

【解決手段】

ＤＴＤ＋パターンツリー作成部５１は、ＤＴＤ＋パターン情報Ｒが定義する各要素の階層構造をツリーとして表すとともに、ツリーの各ノードに、該当する要素に関して指定された記述パターンを付加する。全体コントロール部５２は、このツリーの各ノード毎に、指定された記述パターンに従った検索をパターン検索部５３に依頼する。パターン検索部５３は、指定された記述パターンに合致した領域を処理対象文書Ｔから抽出して、全体コントロール部５２に回答する。全体コントロール部５２は、各要素に対応するものとして回答されたテキストの領域の前後に、その要素に対応したタグを付加することにより、構造化文書Ｏを出力する。

【選択図】　図３

認定・付加情報

特許出願の番号	特願２０００－０２７４６０
受付番号	５００００１２５１２９
書類名	特許願
担当官	第七担当上席　　００９６
作成日	平成１２年　２月　７日

＜認定情報・付加情報＞

【提出日】　平成12年 2月 4日

次頁無

出　願　人　履　歴　情　報

識別番号　　〔０００００５２２３〕

１．変更年月日　　１９９６年　３月２６日
［変更理由］　住所変更
住　所　神奈川県川崎市中原区上小田中４丁目１番１号
氏　名　富士通株式会社